NPCC
National Police Chiefs' Council

POLICE
DIGITAL
SERVICE

CYBER STANDARD DOCUMENT

ARTIFICIAL INTELLIGENCE

ABSTRACT:

This standard brings together a set of control requirements for the use of Artificial Intelligence (AI) in policing. To help the reader in this new area, Artificial Intelligence has been defined, along with a number of its sub-categories. This standard has an additional section targeted at developers and data scientists, to provide more detailed guidance, when developing AI-based solutions.

This standard adheres to the National Policing Community Security Policy Framework and is a suitable reference for community members, notably those who build and implement IT systems on behalf of national policing.

ISSUED	September 2023
PLANNED REVIEW DATE	September 2024
DISTRIBUTION	Community Security Policy Framework Members
STANDARD VALIDITY STATEMENT This document is due for review on the date shown above. After this date, the document may become invalid. Members should ensure that they are consulting the currently valid version of the documentation.	

Document Information

Document Location

PDS - National Policing Policies & Standards

Revision History

Version	Author	Description	Date
0.1	Dean Noble	Initial draft version	01/09/2023
0.2	Dean Noble	Updated after feedback from Head of Audit, Risk & Compliance	14/09/2023
0.3	Dean Noble	Updated after NCSC, Cyber Architecture and NCPSWG feedback	30/09/2023

Approvals

Version	Name	Role	Date
1.0	NCPSB	National Cyber Policy & Standards Board	04/10/23

Document References

Document Name	Version	Date
ISF - Standard of Good Practice (for Information Security)	v2022	07/2022
ISO 27002:2022 - Information security, Cybersecurity and privacy protection – Information security controls	v2022	02/2022
CIS Controls	v8	05/2021
NIST Cyber Security Framework	v1.1	04/2018
CSA Cloud Controls Matrix	v4	01/2021
10 Steps to Cyber Security - NCSC.GOV.UK	Web Doc	05/2021
Artificial Intelligence Toolkit (interpol.int)	-	06/2023
ChatGPT-Impacts on Law Enforcement- August 2023.pdf	-	04/2023
Interpol - Principles for Responsible AI Innovation	-	06/2023
A guide to using artificial intelligence in the public sector - GOV.UK (www.gov.uk)	Web Doc	06/2019
AI Glossary: Artificial intelligence, in so many words | Science	-	07/2017
Data Protection Act 2018 - GOV.UK (www.gov.uk)	-	05/2018
A guide to the data protection principles | ICO	Web Doc	05/2018
The Technology Code of Practice - GOV.UK (www.gov.uk)	Web Doc	07/2021
OWASP Top 10 for Large Language Model Applications | OWASP Foundation	v1.0.1	08/2023
ISF - Securing Operational Machine Learning Systems	-	06/2023
Intelligent security tools - NCSC.GOV.UK	Web Doc	04/2019
NPCC - Principles for Using Artificial Intelligence (AI) in Policing	Web Doc	07/2023
Principles for the security of machine learning - NCSC.GOV.UK	1.0	08/2022

Contents

Document Information 3
Document Location 3
Revision History 3
Approvals 3
Document References 4
Community Security Policy Commitment 6
Introduction 6
Definitions 7
Owner 8
Purpose 8
Audience 9
Scope 9
Requirements 10
Use of LLM Tools, e.g. ChatGPT, Google Bard, Meta LLaMA 11
Communication Approach 27
Review Cycle 27
Document Compliance Requirements 27
Equality Impact Assessment 27
Appendix A – Terms and Abbreviations 28

Community Security Policy Commitment

National Policing and its community members recognise that threats to policing information assets present significant risk to policing operations. National policing and its community members are committed to managing information security and risk and maintaining an appropriate response to current and emerging threats, as an enabling mechanism for policing to achieve its operational objectives whilst preserving life, property, and civil liberties.

This standard in conjunction with the National Policing Community Security Policy Framework and associated documents sets out national policing requirements for the use of Artificial Intelligence in Policing.

Introduction

In recent years we have seen Artificial Intelligence (AI) technologies become embedded in society and our daily lives.

AI technologies have huge potential to support the work of the policing community. AI is already in use across policing and there is no doubt there are many more opportunities for its use.

At the same time, current AI systems have limitations and risks that require awareness and careful consideration by the policing community to either avoid or sufficiently mitigate the issues that can result from their use in police work.

With the recent developmental leaps in AI capabilities, particularly around Generative AI, the public debate around legal and ethical implications of AI systems, as well as the negative effects they could have on society and humanity, has exploded. This standard does not seek to address the legal or ethical implications, but these must be considered.

The AI term is used loosely in the both the technology and broader community and now covers a multitude of sub-categories of AI, the definitions in this document try to explain some of the more common categories, however whichever type of AI is in use, the following is common, software security relies on understanding how a component or system works. This allows a system owner to test for and assess vulnerabilities, which can then be mitigated or accepted. AI and particularly the category referred to as Machine Learning (ML), present particular challenges, because as the name implies, the system learns how to do things itself, and therefore the owner will likely be unable to interpret the logic and understand why the system is doing what it is doing. To quote NCSC[1]

'To summarise, what we are really asking ….

[1] Introducing our new machine learning security principles - NCSC.GOV.UK

How confident would you be proposing, or agreeing to use, a product that you know has vulnerabilities inherent to the product type:

- *for which you don't truly understand the logic*
- *that you can't comprehensively test*
- *and which, once in operation, you're going to allow your users to affect its logic*

Oh, and to add to this . . . it may well contain a representative format of the (possibly sensitive) data on which it was trained.'

As the opening paragraphs state AI is already in use across policing, some of that use will have been planned, but with the huge take-up for tools such as ChatGPT, which have been integrated into browsers, some use may not be planned and may be uncontrolled. It is important that policing takes the recommendations from this standard and begins to apply them across their organisations.

Definitions

There are many definitions available for Artificial Intelligence (AI). For the purposes of this document, we have adopted and expanded on the definition provided by the NPCC endorsed – 'Principles for Using Artificial Intelligence (AI) in Policing', written by Science & Technology in Policing. While we reference AI throughout this document, it is intended that reference to AI, covers all of the below.

What is Artificial Intelligence? [2]

There is no definitive definition of Artificial Intelligence (Alan Turing Institute, 2021), and AI is often used to refer to related applications such as automation, neural networks, and machine learning. To bring clarity for policing, we adopt the following definitions:

- **Artificial intelligence (AI)** refers to a machine that learns, generalizes, or infers meaning from input, thereby reproducing or surpassing human performance. An example is using image analysis to determine whether a video contains sexual activity with a child. The term AI can also be used loosely to describe a machine's ability to perform repetitive tasks without guidance.
- **Machine learning (ML)** refers to algorithms that leverage new data to improve their ability to make predictions or decisions, without having been explicitly programmed to do so. ML is a widely used form of AI that has contributed to innovations such as speech recognition and fraud detection.
- **Advanced Data Analytics (ADA)** uses subject matter expertise and techniques that are typically beyond those of traditional business intelligence to extract insights and make recommendations from complex data. The techniques vary widely, from data visualisation to complex linear models to

[2] NPCC - Principles for Using Artificial Intelligence (AI) in Policing

language analytics. An example is the use of Risk Terrain Modelling to quantify environmental factors that shape risk mapping and resource deployments.

There is other related AI terminology, which start to overlap with the above, but are included here for completeness. The above and below are considered the most common, but there are others. The additional definitions are:

- **Generative Artificial Intelligence (GAI)** is artificial intelligence capable of generating text, images, or other media, using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.[3]
- **Large Language Models (LLMs)** are a subset of GAI, where an algorithm has been trained on a large amount of text-based data, typically scraped from the open internet, and so covers web pages and - depending on the LLM - other sources such as scientific research, books, or social media posts.[4] Examples include ChatGPT, Google Bard and Meta’s LLaMA.
- **Natural Language Processing** is a computer's attempt to “understand” spoken or written language. It must parse vocabulary, grammar, and intent, and allow for variation in language use. The process often involves machine learning.[5]

Owner

National Chief Information Security Officer (NCISO).

Purpose

This standard should empower policing to leverage artificial intelligence responsibly.

It is intended to help policing organisations meet Community Security Policy requirements in a relatively new and fast evolving technology area.

It is important that policing can be innovative with this technology, and so this standard seeks to provide the guardrails, so that innovation can be carried out safely and securely and not putting policing at unnecessary risk of losing public trust and confidence through consequential data loss or loss of policing services.

[3] Generative artificial intelligence - Wikipedia

[4] ChatGPT and LLMs: what's the risk - NCSC.GOV.UK

[5] AI Glossary: Artificial intelligence, in so many words | Science

Audience

Members of the Policing Community of Trust.

More specifically the standard is targeted at, architects, developers, data scientists and security experts tasked with designing and building solutions, applications and plugins leveraging AI related technologies.

The following should also be aware of the content of this standard, in order that they can provide appropriate oversight and governance of the use of AI related technologies within policing:

- Senior Information Risk Owners (SIROs)
- Information Asset Owners (IAOs)
- Information & Cyber risk practitioners and managers
- Auditors providing assurance services to PDS or policing.

Finally, Policing's reliance on third parties means that suppliers acting as service providers or developing products or services for PDS or policing, should also be made aware of and comply with the content of this standard, in relation to their work on Policing systems and data.

Scope

In scope for this standard are cyber security considerations and requirements for the:

- Acquisition and implementation of solutions that incorporate AI.
- Development of solutions with integrated AI.
- Use of AI tools, e.g. ChatGPT, Google Bard and LLaMA.

Out of scope:

- Ethical considerations – Instead please reference Understanding artificial intelligence ethics and safety - GOV.UK (www.gov.uk) and/or Data Ethics Framework (publishing.service.gov.uk)
- Legal or Regulatory considerations, other than those directly related to cyber.
- Weaponization and/or use of AI against policing organisations, e.g. AI-generated phishing, vishing, fake profiles, malicious chatbots and advanced malware.
- Other key AI usage principles, which are covered by the NPCC endorsed - Principles for Using Artificial Intelligence (AI) in Policing and other relevant government documentation referenced at the beginning of this standard.

Requirements

This section details the minimum requirements for the acquisition, development and use of AI to protect policing from the loss of confidentiality, integrity or availability of the data or loss of availability of the systems and services it relies upon to meet policing outcomes.

The newness of AI systems and their potential for driving dramatic change or unearthing crucial business insights suggests that these systems demand an entirely fresh set of control requirements to monitor, police and curate them. This is not necessarily the case. These systems in many ways are similar to existing technologies in use across policing organisations today, making them manageable in the same way as those familiar systems.

Most control requirements relating to the acquisition and use of AI and AI-based solutions can be found in the existing standards which have been written, or are being written to support the National Policing Community Security Policy and Principles, and will apply to AI tools and AI integrated tools, just as they do to any other digital solution, however AI tools and AI integrated solutions can present a unique set of vulnerabilities, and so the requirements below are to address these.

These requirements are not intended to replace those in other standards and while there will be duplication, they are documented here, as they are particularly pertinent to the acquisition, development, and use of AI-based systems.

This standard and these requirements do not standalone; it is important that all cyber security standards are considered during the acquisition and development of AI-based solutions.

Example existing standards that should be consulted are:

- System Access
- System Development
- Threat & Incident Management
- Cryptography
- Third Party Assurance for Policing

Other standards that will also need to be consulted, but are in development or not yet written, are listed below. Until these are available, Policing organisations should consult their existing equivalent policies and standards:

- Application Management
- System Management
- Networks & Communications
- Technical Security Management
- Business Continuity
- Information Assurance

The above is not an exhaustive list but will provide a solid base when developing AI-based solutions.

This standard includes an additional control requirements section, which is targeted at Developers and Data Scientists and should be considered by Security Professionals working with and assuring AI-based solutions. They are of a more technical nature and not specifically aligned to typical NIST/ISO control frameworks. They have been directly sourced from OWASP recommendations and are correlated to the top 10 known vulnerabilities, when developing AI-based solutions, specifically of the LLM category.

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
1	**Use of LLM Tools, e.g. ChatGPT, Google Bard, Meta LLaMA**		
1.1	When using LLM tools, e.g. ChatGPT for law enforcement purposes, policing organisations should exercise caution regarding potential data confidentiality and disclosure issues. Consideration should be given to any risks of using the solution, and sign-off from the appropriate risk owner.	**NIST CSF:** ID.AM.3&5/ PR.DS.1&2&3/ PR.IP.6/ID.GV.3/ ID.SC.3 **ISO 27002:2022:** 5.09/5.10/5.12/ 5.13/5.14/5.32 5.33/5.34/8.10/ 8.11 **ISF SOGP:** IM1.1/1.2/1.3/1.4	Business Impact Assessments, Risk registers, Risk Mitigation Plans, Information Risk Management Framework
1.2	LLM platforms like ChatGPT must not be used for processing police data classified at ‘official’ or above, without prior consultation with the appropriate information security team, a full understanding of the risks, and sign-off from the appropriate risk owner.		
1.3	LLM platforms like ChatGPT must not be used to support operational decision making or in the creation of documents that may enter the Criminal Justice system, without prior consultation with the appropriate operational process owners, legal and information security team, a full understanding of the risks, and sign-off from the appropriate risk owner.	None	

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
1.4	To minimise the risk of unintentional breaches of the above requirements, policing organisations should consider blocking access to online LLM tools and setting up an exception process for allowing controlled access to LLM tools, where there is a valid need and a suitable risk assessment has been carried out, and any residual risks accepted by the appropriate risk owner.	**NIST CSF:** DE.CM.1 **ISO 27002:2022:** 8.09/8.21/8.22 /8.23 **ISF SOGP:** NC1.5	Perimeter defence technology configurations, e.g. Secure Web Gateways (SWG), Business Impact Assessments, Risk registers, Risk Mitigation Plans, Information Risk Management Framework, and third-party assurance reports.
1.5	Use network monitoring and end point management systems to spot when employees visit sites offering LLM-type services or apps. Steer these workers towards approved alternatives and advise about appropriate use.	**NIST CSF:** PR.IP.2/DE.AE.3&4&5 /DE.CM.1&7/DE.DP.2 /DE.DP.4&5/RS.AN.1 **ISO 27002:2022:** 5.25/8.15/8.16 **ISF SOGP:** TM1.3	Network monitoring and end point management systems LLDs and reports
1.6	Any access to and upload of data to online AI tools should be logged, monitored and alerted if not approved to the local Data Protection Team, and appropriate action taken.	**NIST CSF:** ID.AM.3/PR.DS.1/ PR.DS.5 **ISO 27002:2022:** 5.12/8.12 **ISF SOGP:** IM1.5	SWG & DLP LLDs and reports
1.7	Educate all staff and officers about the risks of AI tools, such as ChatGPT and the appropriate use of these tools, where they are authorised to use them.	**NIST CSF:** PR.AT.1&2/ PR.IP.11 **ISO 27002:2022:** 6.03/7.07 **ISF SOGP:** PM2.1/PM2.2	Education and Awareness (E&A) materials relating to the use of AI and E&A tracking and measurement of

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
1.8	Educate IT support staff about the risks of AI tools, such as ChatGPT and their appropriate use and coach them in giving advice to anyone seeking help on how to use these tools.	**NIST CSF:** PR.AT.1to3 **ISO 27002:2022:** 6.03 **ISF SOGP:** PM2.3	effectiveness of E&A campaign
2	**General**		
2.1	Before embarking on AI-based solution development or acquisition, policing organisations must ensure they have the requisite skilled people, process and technology elements required to set up, run and maintain operational AI systems and cope with their outputs.	**NIST CSF:** PR.DS.4/PR.MA.1 PR.PT.4&5/PR.IP.1 DE.AE.1/DE.CM.7 ID.AM.6/ID.GV.2 PR.AT.2&5 **ISO 27002:2022:** 8.17 / 5.02 **ISF SOGP:** SY1.1 / SM2.1	Project Plans, Target Operating Models, Role Descriptions, Skills Matrices, Process Documentation and Technology
2.2	To ensure the above is in place and maintained and other control requirements listed here are implemented and adhered to, policing organisations should set up a suitable AI governance framework or incorporate it into a suitable existing governance framework.	**NIST CSF:** ID.BE.3/ID.GV.4 ID.RM.3/PR.IP.7 **ISO 27002:2022:** N/A **ISF SOGP:** SG1.1	Governance ToRs, Governance Meeting minutes and action and decision logs
2.3	Work with legal, data protection and privacy teams to maintain awareness of the evolving legal and regulatory changes impacting the use of and security of AI, on an ongoing basis.	**NIST CSF:** ID.BE.1&2 ID.GV.3 **ISO 27002:2022:** 5.31/5.32/5.33/ 5.35/5.36/8.24 **ISF SOGP:** SM2.5	Relevant legislation and regulation are at hand

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.4	Monitor guidance from regulators on their standards for datasets. Test and sample datasets to expose systemic, computational or human-cognitive bias, to minimise the risk of the AI solution basing decisions on bias data.	**NIST CSF:** ID.BE.1&2/ID.GV.3 /PR.DS.7 **ISO 27002:2022:** 5.31/5.32/5.33/ 5.35/5.36/8.24/ 8.29/8.33 **ISF SOGP:** SM2.5/SD2.5	Regulatory standards available, Testing Reports
2.5	AI solutions expose little information about how decisions are reached, and organisations could be exposed to regulatory sanctions, if explanations or justifications are absent for decisions made. Therefore policing organisations should: • Test extensively to gain as much understanding as possible about what outputs are likely. Test against different user and customer groups to probe for bias. • Run parallel models to explore how decisions and outputs can change. • Talk to regulators about their expectations to set levels for transparency and explainability. • Prove and demonstrate the efficacy of the entire system. If the AI solution has been acquired, then the policing organisation should seek clarification from the provider that the above has been carried out.	**NIST CSF:** N/A **ISO 27002:2022:** 5.35/8.16 **ISF SOGP:** AS1.1	Test reports, Third Party Assurance reports

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.6	Adopt a risk-based, rather than compliance-centric, approach to help local information security teams and other appropriate stakeholders meet regulatory and legal requirements, but not stifle innovation and the potential benefits of the use of AI in policing.	**NIST CSF:** ID.GV.4/ID.RA.6 ID.RM.1/PRIP.7 **ISO 27002:2022:** N/A **ISF SOGP:** IR1.1	Business Impact Assessments, Risk registers, Risk Mitigation Plans, Information Risk Management Framework
2.7	Ensure regular policy/standard gap analysis is undertaken to address policy/standard shortcomings as AI technology evolves. Update and socialise existing policy/standards, particularly on acceptable use, data leakage prevention (DLP) and data handling, to emphasise what is relevant and appropriate when using AI systems.	**NIST CSF:** ID.AM.3/PR.DS.1/ PR.DS.5/ID.GV.1/ RS.CO.2 **ISO 27002:2022:** 5.01/5.12/6.04/8.12 **ISF SOGP:** IM1.5 / SM1.1	Up to date cyber security policy and standards relating to AI
2.8	Review Incident Response Plans and ensure potential AI related incidents are adequately covered and then tested, e.g. personally identifiable data being leaked.	**NIST CSF:** PR.IP.2/ DE.AE.2&3&4&5 /DE.CM.1&6&7 /DE.DP.2&4&5 /RS.AN.1 **ISO 27002:2022:** 5.25/8.15/8.16 **ISF SOGP:** TM1.3	Up to date IRP, IRP Test Plans, IRP Test results, IRP improvement plans
2.9	When embarking on an AI related project, consider whether AI is necessary, based on the problem you are trying to solve and then follow a 'Secure by Design' (SbD) methodology and ensure a Business Impact Assessment is conducted at the outset.	**NIST CSF:** ID.RA.4 **ISO 27002:2022:** N/A **ISF SOGP:** IR2.2	SbD process documentation, BIAs, SbD project artefacts

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.10	Develop logging and monitoring requirements for each AI instance and suitable playbooks for response to AI related security events.	**NIST CSF:** PR.PT.1/DE.AE.1&3/ DE.CM.1&3&7/ DE.DP.2&4/RS.CO.2 /RS.AN.1 **ISO 27002:2022:** 8.15/8.17 **ISF SOGP:** TM1.2	Design decision log, Low Level Design documents, Security Operations Centre Playbooks
2.11	Put access controls in place to limit who can input or amend data within the AI solution, and alert when unauthorised or unexpected changes are made.	**NIST CSF:** PR.AC.1&4&6/ PR.PT.3 **ISO 27002:2022:** 5.15 **ISF SOGP:** SA1.1	Design decision log, Low Level Design documents
2.12	Use vulnerability management tools and techniques to identify at risk AI systems and harden them against compromise.	**NIST CSF:** ID.RA.1&2/PR.IP.12 /DE.CM.8/RS.AN.5/ RS.MI.3 **ISO 27002:2022:** 8.08/8.18 **ISF SOGP:** TM1.1	VM processes, VM tools, VM remediation logs
2.13	Update and maintain threat intelligence feeds to watch for attacks or compromises of AI tools.	NIST CSF: ID.RA.2&3/DE.AE.2 /RS.CO.5/RS.AN.5 **ISO 27002:2022:** 5.07 **ISF SOGP:** TM1.4	Threat Intel monitoring list and threat intel reports relating to AI
2.14	Check whether suppliers or open sources meet existing security standards on ID management, access control and authentication. Ensure supply chain assurance extends to software and service providers who help to develop and maintain organisational AI systems.	**NIST CSF:** ID.SC.4 **ISO 27002:2022:** 5.19/5.21/5.22 **ISF SOGP:** SC1.4	Third Party Assurance Reports

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.15	Clarify business impact of AI disruption (e.g. outages) and design for the appropriate levels of uptime and resiliency. Ensure the robustness of technology platforms underlying the AI system, meet the levels of uptime and resiliency required.	**NIST CSF:** PR.AC.4/ID.BE.5/ PR.IP.5/PR.PT.5 **ISO 27002:2022:** 5.3/7.11/8.14/8.27 **ISF SOGP:** TS1.1 / BC1.3	Business Impact Assessment, Low Level Designs and Design Decision Logs
2.16	To minimise service disruption to the AI solution, policing organisations should ensure: • Appropriate levels of network segregation and access control are in place to limit access to the AI system via lateral movement and direct access. • Use threat intelligence and modelling techniques to gauge current level of threat to the AI solution. • Use red team testing techniques to explore possible attack patterns. • Manage access rights. • Manage how much information the AI model returns to queries to limit the ability of threat actors to gather useful attack data.	**NIST CSF:** PR.IP.1/ PR.AC.1&4&5&6 /PR.PT.3&4/ ID.RA.2&3/DE.AE.2 /RS.CO.5/RS.AN.5/ DE.AE.3/ DE.CM.1&3&3&7/ RS.RP.1/RS.IM.1 **ISO 27002:2022:** 8.09/8.20/5.07 /5.15 **ISF SOGP:** NC1.1/TM1.4/ TM1.5/SA1.1	Business Impact Assessment, Low Level Designs, Design Decision Logs, Threat Intelligence Reports, Red Team Reports,

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.17	To prevent potential copyright claims being made against policing, for using data or sources without permission or an appropriate licence (a particular issue related to open-source AI models), policing organisations should: • Investigate data sources used to build an AI model, especially when procuring AI services. Gain assurance from suppliers on sources and their use of copyrighted material, in particular when deploying an open-source model. • Engage with your local legal and regulatory teams to obtain the latest guidance on IP and how it is used to build AI models.	**NIST CSF:** ID.SC.2/ID.BE.1&2 /ID.GV.3 **ISO 27002:2022:** 5.31/5.32/5.33/ 5.35/5.36/8.24/ 5.19/8.3 **ISF SOGP:** SC1.2 / SM2.5	Third Party Assurance Reports, Legal/Regulatory guidance obtained
2.18	As another technology asset, it must be recorded in the relevant asset inventory, with a clear description of its approved use, who owns it and other data relevant to the AI system and its use.	**NIST CSF:** ID.AM.1&2/PR.DS.3 **ISO 27002:2022:** 5.09/5.10 **ISF SOGP:** SM2.6/SD2.3	Technology Asset Inventory
2.19	Communication of this standard or a suitable summary to all potential users, developers or implementors of AI tooling.	**NIST CSF:** PR.AT.1/ID.GV.1/ **ISO 27002:2022:** 6.03/5.01 **ISF SOGP:** PM2.1/SM1.1/SM1.2	Communication Plan & Communication Artefacts

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
2.20	Carry-out an AI discovery exercise, i.e. who is already using AI, or in the process of developing or acquiring AI solutions and for what purpose. Results from the above exercise should be: • Added to the aforementioned technology asset register. • Risk assessed against the organisation's information risk management framework. • Any previously unknown risks identified recorded, remediated and/or accepted by an appropriate risk owner.	**NIST CSF:** ID.AM.1/PR.DS.3 **ISO 27002:2022:** 5.09/5.10 **ISF SOGP:** SM2.6	Discovery Plan, Communication Artefacts, Response Register, Technology Asset Register, Risk Assessment Reports & Risk Register
3.0	**Acquisition and Use of AI-based Cyber Security Specific Tools**		
3.1	Ensure that data sources used to inform security decisions made by the tool have known origin and are accurate, i.e. garbage in, garbage out.	**NIST CSF:** ID.AM.3&5/ PR.DS.1,2,3&6/ PR.IP.6/ID.SC.3 **ISO 27002:2022:** 5.09/5.12/5.13/ 5.14/5.33/8.10/ 8.11/8.29 **ISF SOGP:** IM1.1/PM1.2/BA1.3	Business Impact Assessments, Risk registers, Third Party Assurance Reports.
3.2	Develop an assurance regime that regularly checks the data origins and who supplies that data.	**NIST CSF:** N/A **ISO 27002:2022:** 5.35/8.16 **ISF SOGP:** AS1.1	Business Impact Assessments, Risk registers, Third Party Assurance Reports.
3.3	Implement assurance of third parties supplying data to detect problematic sources and potential for bias at the earliest opportunity.	**NIST CSF:** ID.AM.6/ID.BE.1/ ID.GV.2/ID.SC.1to5 **ISO 27002:2022:** 5.19/5.21/5.22 **ISF SOGP:** SC1.1	Third Party Assurance Reports

Reference	Minimum requirement	Control reference	Compliance Metric / Artefacts
3.4	Data sources can be actively attacked to undermine their usefulness. Therefore policing organisations should: • Regularly sample datasets to spot anomalies. • Establish an early warning system that alerts when unexpected data types or results appear. • Scrutinise the supply chain for gaps that attackers can insert themselves into. • Regularly assure connections with external suppliers for any opportunities for model loss, damage or pollution of the dataset.	None	Evidence of data sampling, Low Level Designs, supply chain analysis, ITHC/Pen Test reports
3.5	Policing cyber security teams should not become over-reliant on the decisions of an AI based security system and therefore should: • Establish guardrails or dashboards for what normal security operation looks like. Regularly audit to gauge whether the system is operating as expected. • Establish a rapid reaction process to examine and investigate exceptions and anomalies to determine causes. • Keep humans 'in the loop' – build processes and procedures around the AI based security solutions, that ensure there are suitable human checkpoints.	**NIST CSF:** DE.AE.2to4/RS.RP.1 /RS.CO.1to5/ RS.AN.2to4/RS.MI.1to2 /RS.CP.1/RC.IM.1 /ID.BE.3/ID.GV.4 /ID.RM.3 **ISO 27002:2022:** 5.24/5.25/5.26/ 5.27/6.08 **ISF SOGP:** AS1.3/TM2.2 /SG1.1	Documented guardrails, dashboards, audit reports, response process, processes and procedures

The following control requirements are more technical in nature and targeted at Developers and Data Scientists who may be developing with AI tools, typically LLMs at present.

Checking these controls are in place will typically be carried out by a Security Professional, who would be best placed to work alongside the Developers and Data Scientists while they are developing in this space.

These control requirements have been taken from the 'OWASP Top 10 for LLMs 2023 v1.0.1' (issued August 2023). Updates to this document should be monitored and reflected in these control requirements, as this is currently a fast-developing area.

Reference	Control Requirement
4.0	***To reduce the vulnerability of 'Prompt Injections'***
4.1	Enforce privilege control on LLM access to backend systems. Provide the LLM with its own API tokens for extensible functionality, such as plugins, data access, and function-level permissions. Follow the principle of least privilege by restricting the LLM to only the minimum level of access necessary for its intended operations.
4.2	Implement human in the loop for extensible functionality. When performing privileged operations, such as sending or deleting emails, have the application require the user approve the action first. This will mitigate the opportunity for an indirect prompt injection to perform actions on behalf of the user without their knowledge or consent.
4.3	Segregate external content from user prompts. separate and denote where untrusted content is being used to limit their influence on user prompts. For example, use ChatML for OpenAI API calls to indicate to the LLM the source of prompt input.
4.4	Establish trust boundaries between the LLM, external sources, and extensible functionality (e.g., plugins or downstream functions). Treat the LLM as an untrusted user and maintain final user control on decision-making processes. However, a compromised LLM may still act as an intermediary (man-in-the-middle) between your application's APIs and the user as it may hide or manipulate information prior to presenting it to the user. Highlight potentially untrustworthy responses visually to the user.
5.0	***To reduce the vulnerability of 'Insecure Output Handling'***
5.1	Apply input validation on responses coming from the model to backend functions. Follow the OWASP ASVS (Application Security Verification Standard) guidelines to ensure effective input validation and sanitization.
5.2	Encode model output back to users to mitigate undesired code execution by JavaScript or Markdown. OWASP ASVS provides detailed guidance on output encoding.
6.0	***To reduce the vulnerability of 'Training Data Poisoning'***
6.1	Verify the supply chain of the training data, especially, when sourced externally as well as maintaining attestations, similar to the "SBOM" (Software Bill Of Materials) methodology.
6.2	Verify the correct legitimacy of targeted data sources and data obtained during both the training and fine-tuning stages.
6.3	Verify your use-case for the LLM and the application it will integrate to. Craft different models via separate training data or fine-tuning for different use-cases to create a more granular and accurate generative AI Output as per it's defined use-case.
6.4	Ensure sufficient sandboxing is present to prevent the model from scraping unintended data sources which could hinder the machine learning output.
6.5	Use strict vetting or input filters for specific training data or categories of data sources to control volume of falsified data. Data sanitisation, with techniques such as

	Statistical outlier detection and anomaly detection methods to detect and remove adversarial data from potentially being fed into the fine-tuning process.
6.6	Adversarial robustness techniques such as federated learning and constraints to minimize the effect of outliers or adversarial training to be vigorous against worst-case perturbations of the training data. a. An "MLSecOps" approach could be to include adversarial robustness to the training lifecycle with the auto poisoning technique. b. An example repository of this would be Autopoison testing, including both attacks such as Content Injection Attacks ("how to inject your brand into the LLMs responses") and Refusal Attacks ("always making the model refuse to respond") that can be accomplished with this approach.
6.7	Testing and Detection, by measuring the loss during the training stage and analysing trained models to detect signs of a poisoning attack by analysing model behaviour on specific test inputs. a. Monitoring and alerting on number of skewed responses exceeding a threshold. b. Use of a human loop to review responses and auditing. c. Implement dedicated LLMs to benchmark against undesired consequences and train other LLMs using reinforcement learning techniques. d. Perform LLM-based red team exercises or LLM vulnerability scanning in the testing phases of the LLM lifecycle.
7.0	***To minimise the risk of 'Model Denial of Service'***
7.1	Implement input validation and sanitisation to ensure user input adheres to defined limits and filters out any malicious content.
7.2	Cap resources use per request or step, so that requests involving complex parts execute more slowly.
7.3	Enforce API rate limits to restrict the number of requests an individual user or IP address can make within a specific timeframe.
7.4	Limit the number of queued actions and the number of total actions in a system reacting to LLM responses.
7.5	Continuously monitor the resource utilisation of the LLM to identify abnormal spikes or patterns that may indicate a DoS attack.
7.6	Set strict input limits based on the LLMs context window to prevent overload and resource exhaustion.
7.7	Promote awareness amongst developers about potential DoS vulnerabilities in LLMs and provide guidelines for secure LLM implementation.
8.0	***To minimise the risk of 'Supply Chain Vulnerabilities'***
8.1	Carefully vet data sources and suppliers, including T&Cs and their privacy policies, only using trusted suppliers. Ensure adequate and independently audited security is in place and that model operator policies align with your data protection policies, i.e. your data is not used for training their models; similarly, seek assurances and legal mitigations against using copyrighted material from model maintainers.

8.2	Only use reputable plug-ins and ensure they have been tested for your application requirements. LLM-Insecure Plugin Design provides information on the LLM-aspects of Insecure Plugin design you should test against to mitigate risks from using third-party plugins.
8.3	Understand and apply the mitigations found in the OWASP Top Ten's 'A06:2021 – Vulnerable and Outdated Components'. This includes vulnerability scanning, management, and patching components. For development environments with access to sensitive data, apply these controls in those environments, too.
8.4	Maintain an up-to-date inventory of components using a Software Bill of Materials (SBOM) to ensure you have an up-to-date, accurate, and signed inventory preventing tampering with deployed packages. SBOMs can be used to detect and alert for new, zero-day vulnerabilities quickly.
8.5	At the time of writing, SBOMs do not cover models, their artifacts, and datasets; If your LLM application uses its own model, you should use MLOps best practices and platforms offering secure model repositories with data, model, and experiment tracking.
8.6	You should also use model and code signing when using external models and suppliers.
8.7	Anomaly detection and adversarial robustness tests on supplied models and data can help detect tampering and poisoning as discussed in '*To minimise the risk of Training Data Poisoning'* above; Ideally, this should be part of MLOps pipelines; however, these are emerging techniques and may be easier implemented as part of red teaming exercises.
8.8	Implement sufficient monitoring to cover component and environment vulnerabilities scanning, use of unauthorised plugins, and out-of-date components, Including the model and its artifacts.
8.9	Implement a patching policy to mitigate vulnerable or outdated components. Ensure that the application relies on a maintained version of APIs and the underlying model.
8.10	Regularly review and audit supplier Security and Access, ensuring no changes in their security posture or T&Cs.
9.0	***To minimise the risk of 'Sensitive Information Disclosure'***
9.1	Integrate adequate data sanitisation and scrubbing techniques to prevent user data from entering the training model data.
9.2	Implement robust input validation and sanitization methods to identify and filter out potential malicious inputs to prevent the model from being poisoned.
9.3	When enriching the model with data and if fine-tuning a model: (I.e. data fed into the model before or during deployment) a. Anything that is deemed sensitive in the fine-tuning data has the potential to be revealed to a user. Therefore, apply the rule of least privilege and do not train the model on information that the highest-privileged user can access which may be displayed to a lower-privileged user. b. Access to external data sources (orchestration of data at runtime) should be limited. c. Apply strict access control methods to external data sources and a rigorous approach to maintaining a secure supply chain.
10.0	***To reduce the vulnerability of 'Insecure Plugin Design'***

10.1	Plugins should enforce strict parameterized input wherever possible and include type and range checks on inputs. When this is not possible, a second layer of typed calls should be introduced, parsing requests and applying validation and sanitization. When freeform input must be accepted because of application semantics, it should be carefully inspected to ensure that no potentially harmful methods are being called.
10.2	Plugin developers should apply OWASP's recommendations in ASVS (Application Security Verification Standard[6]) to ensure effective input validation and sanitization.
10.3	Plugins should be inspected and tested thoroughly to ensure adequate validation. Use Static Application Security Testing (SAST) scans as well as Dynamic and Interactive Application Testing (DAST, IAST) in development pipelines.
10.4	Plugins should be designed to minimise the impact of any insecure input parameter exploitation following the OWASP ASVS Access Control Guidelines[7]. This includes least-privilege access control, exposing as little functionality as possible while still performing its desired function.
10.5	Plugins should use appropriate authentication identities, such as OAuth2, to apply effective authorisation and access control. Additionally, API keys should be used to provide context for custom authorisation decisions which reflect the plugin route rather than the default interactive user.
10.6	Require manual user authorisation and confirmation of any action taken by sensitive plugins.
10.7	Plugins are typically REST APIs, so developers should apply the recommendations found in OWASP Top 10 API Security Risks – 2023[8] to minimise generic vulnerabilities.
11.0	***To reduce the vulnerability of 'Excessive Agency'***
11.1	Limit plugins/tools that LLM agents are allowed to call to only the minimum functions necessary. For example, if an LLM-based system does not require the ability to fetch the contents of a URL, then such a plugin should not be offered to the LLM agent.
11.2	Limit the functions that are implemented in LLM plugins/tools to the minimum necessary. For example, a plugin that accesses a user's mailbox to summarise emails may only require the ability to read emails, so the plugin should not contain other functionality such as deleting or sending messages.
11.3	Avoid open-ended functions where possible (e.g. run a shell command, fetch a URL, etc.) and use plugins/tools with more granular functionality. For example, an LLM-based app may need to write some output to a file. If this were implemented using a plugin to run a shell function, then the scope for undesirable actions is very large (any other shell commands could be executed). A more secure alternative would be to build a file-writing plugin that could only support that specific functionality.
11.4	Limit the permissions that LLM plugins/tools are granted to other systems to the minimum necessary in order to limit the scope of undesirable actions. For example, an LLM agent that uses a product database in order to make purchase recommendations to a customer might only need read access to a 'products' table; it should not have access to other tables, nor the

[6] OWASP Application Security Verification Standard | OWASP Foundation

[7] OWASP Developer Guide | Enforce Access Controls Checklist | OWASP Foundation

[8] OWASP Top 10 API Security Risks – 2023 - OWASP API Security Top 10

	ability to insert, update or delete records. This should be enforced by applying appropriate database permissions for the identity that the LLM plugin uses to connect to the database.
11.5	Track user authorisation and security scope to ensure actions taken on behalf of a user are executed on downstream systems in the context of that specific user, and with the minimum of privileges necessary. For example, an LLM plugin that reads a user's code repo should require the user to authenticate via OAuth and with the minimum scope required.
11.6	Utilise human-in-the-loop control to require a human to approve all actions before they are taken. This may be implemented in a downstream system (outside the scope of the LLM application) or within the LLM plugin/tool itself. For example, an LLM-based app that creates and posts social media content on behalf of a user should include a user approval routine within the plugin/tool/API that implements the 'post' operation.
11.7	Implement authorisation in downstream systems rather than relying on an LLM to decide if an action is allowed or not. When implementing tools/plugins enforce the complete mediation principle so that all requests made to downstream systems via the plugins/tools are validated against security policies.
11.8	Log and monitor the activity of LLM plugins/tools and downstream systems to identify where undesirable actions are taking place and respond accordingly.
11.9	Implement rate-limiting to reduce the number of undesirable actions that can take place within a given time period, increasing the opportunity to discover undesirable actions through monitoring before significant damage can occur.
12.0	***To minimise the risk of 'Overreliance'***
12.1	Regularly monitor and review the LLM outputs. Use self-consistency or voting techniques to filter out inconsistent text. Comparing multiple model responses for a single prompt can better judge the quality and consistency of output.
12.2	Cross-check the LLM output with trusted external sources. This additional layer of validation can help ensure the information provided by the model is accurate and reliable.
12.3	Enhance the model with fine-tuning or embeddings to improve output quality. Generic pre-trained models are more likely to produce inaccurate information compared to tuned models in a particular domain. Techniques such as prompt engineering, parameter efficient tuning (PET), full model tuning, and chain of thought prompting can be employed for this purpose.
12.4	Implement automatic validation mechanisms that can cross-verify the generated output against known facts or data. This can provide an additional layer of security and mitigate the risks associated with hallucinations.
12.5	Break down complex tasks into manageable subtasks and assign them to different agents. This not only helps in managing complexity, but it also reduces the chances of hallucinations as each agent can be held accountable for a smaller task.
12.6	Communicate the risks and limitations associated with using LLMs. This includes potential for information inaccuracies, and other risks. Effective risk communication can prepare users for potential issues and help them make informed decisions.

12.7	Build APIs and user interfaces that encourage responsible and safe use of LLMs. This can involve measures such as contents filters, user warnings about potential inaccuracies, and clear labelling of AI-generated content.
12.8	When using LLMs in development environments, establish secure coding practices and guidelines to prevent the integration of possible vulnerabilities.
13.0	***To minimise the risk of 'Model Theft'***
13.1	Implement strong access controls (e.g., RBAC and rule of least privilege) and strong authentication mechanisms to limit unauthorised access to LLM model repositories and training environments.
13.2	Restrict the LLMs access to network resources, internal services, and APIs.
13.3	Regularly monitor and audit access logs and activities related to LLM model repositories to detect and respond to any suspicious or unauthorized behaviour promptly.
13.4	Automate MLOps deployment with governance and track: ng and approval workflows to tighten access and deployment controls within the infrastructure.
13.5	Implement controls and mitigation strategies to mitigate and/or reduce risk of prompt injection techniques causing side-channel attacks.
13.6	Rate Limiting of API calls where applicable and/or filters to reduce risk of data exfiltration from the LLMs applications, or implement techniques to detect (e.g., DLP) extraction activity from other monitoring systems.
13.7	Implement adversarial robustness training to help detect extraction queries and tighten physical security measures.
13.8	Implement a watermarking framework into the embedding and detection stages of an LLMs lifecycle.

Communication Approach

This standard will be communicated as follows:

1. Internal peer review by the members of the National Cyber Policy and Standard Working Group (NCPSWGG), which includes representatives from PDS and participating forces.
2. Presentation to the National Cyber Policy & Standards Board (NCPSB) for approval.
3. Formal publication and external distribution to PDS community, police forces and associated bodies.

This standard should be distributed within IT and project teams to help complete an initial gap analysis which can inform any implementation plan. This implementation plan can be shared with force SIROs / Security Management Forum / Information Management. Consideration should also be given to raising awareness amongst force personnel of the implementation of this standard where it may affect them.

This standard should be mapped to a project lifecycle and internal governance prior to adoption. Following this, it should be provided to the Information Assurance communities and PMO's and should also be shared with procurement & commercial leads to ensure this is built into procurement activities.

Measurables generated by adopting this standard can also form part of regular Cyber management reporting and audit evidencing.

Review Cycle

NPCC is actively working with government, academia, and industry to fully understand the risks and opportunities around the use of AI in policing. This standard will be kept under review to reflect changes in this fast-evolving technology domain.

As a minimum, this standard will be reviewed at least annually (from the date of publication) and following any major change to Information Assurance (IA) strategy, membership of the community, or an identified major change to the cyber threat landscape. This ensures IA requirements are reviewed and that the standard continues to meet the objectives and strategies of the police service.

Document Compliance Requirements

(Adapt according to Force or PDS Policy needs.)

Equality Impact Assessment

(Adapt according to Force or PDS Policy needs.)

Appendix A – Terms and Abbreviations

Term	Abbreviation	Brief explanation
Adversarial Training	-	Adversarial training is a defensive method to improve the robustness of a model by reducing the malicious effect caused by adversarial attacks.
Advanced Data Analytics	ADA	Uses subject matter expertise and different tools and techniques, e.g. statistical analysis, ML and AI, to extract insights and make recommendations from complex data. An example is the use of Risk Terrain Modelling to quantify environmental factors that shape risk mapping and resource deployments.
Algorithm	-	A set of step-by-step instructions. Computer algorithms can be simple (if it's 3 p.m., send a reminder) or complex (identify pedestrians).[9]
Artificial Intelligence	AI	Refers to a machine that learns, generalizes, or infers meaning from input, thereby reproducing or surpassing human performance. An example is using image analysis to determine whether a video contains sexual activity with a child. The term AI can also be used loosely to describe a machine’s ability to perform repetitive tasks without guidance.
Autopoison Testing	-	Autopoison is an automated data poisoning pipeline. It can be used to deliberately inject poisoned data in to LLMs, to allow developers to understand the impact and improve data quality to defend against poisoning.
Bard	-	Is a conversational generative artificial intelligence chatbot developed by Google, based initially on the LaMDA family of large language models (LLMs) and later the PaLM LLM.[10]
Bidirectional Encoder Representations from Transformers	BERT	Is a family of language models introduced in 2018 by researchers at Google.[11]
Business Impact Assessment	BIA	An assessment of the impact to compromise of confidentiality, integrity, and availability of information assets.

[9] AI Glossary: Artificial intelligence, in so many words | Science

[10] Bard (chatbot) - Wikipedia

[11] BERT (language model) - Wikipedia

Term	Abbreviation	Brief explanation
ChatML	-	ChatML is an NPM package that streamlines context management with OpenAI's chat completion API, making it easier to maintain context throughout AI-powered conversations.
Controls	-	Mitigations or countermeasures to vulnerabilities. These can be technical (a firewall), administrative (policies and procedures) or physical (security guard).
Data Loss Prevention	DLP	Also referred to as Data Leakage Prevention. Typically refers to the process and technology required to prevent sensitive information from being disclosed to unauthorised individuals or systems.
Data Protection Act 2018	DPA	The Data Protection Act 2018 is the UK's implementation of the General Data Protection Regulation (GDPR). Everyone responsible for using personal data has to follow strict rules called 'data protection principles. They must make sure the information is: • used fairly, lawfully and transparently, • used for specified, explicit purposes, • used in a way that is adequate, relevant and limited to only what is necessary, • accurate and, where necessary, kept up to date, • kept for no longer than is necessary, • handled in a way that ensures appropriate security, including protection against unlawful or unauthorised processing, access, loss, destruction or damage.
Deep Learning	-	How a neural network with multiple layers becomes sensitive to progressively more abstract patterns. In parsing a photo, layers might respond first to edges, then paws, then dogs.[12]
Design Decision Tracker	-	This document tracks deviations from a standard and the design decisions made.
Excessive Agency	-	AI-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the AI-based systems.

[12] AI Glossary: Artificial intelligence, in so many words | Science

Term	Abbreviation	Brief explanation
Federated Learning	-	Federated learning (also known as collaborative learning) is a machine learning technique that trains an algorithm via multiple independent sessions, each using its own dataset. This approach stands in contrast to traditional centralized machine learning techniques where local datasets are merged into one training session, as well as to approaches that assume that local data samples are identically distributed.[13]
General Data Protection Regulation	GDPR	The General Data Protection Regulation (Regulation (EU) 2016/679, abbreviated GDPR) is a European Union regulation on Information privacy in the European Union (EU) and the European Economic Area (EEA). The UK's Data Protection Act 2018 is aligned to the GDPR.
Generative Adversarial Networks	-	A pair of jointly trained neural networks that generates realistic new data and improves through competition. One net creates new examples (fake Picassos, say) as the other tries to detect the fakes.[14]
Generative Artificial Intelligence	Generative AI	Is artificial intelligence capable of generating text, images, or other media, using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.[15]
High Level Design	HLD	High-level design (HLD) explains the architecture that would be used to develop a system.
Insecure Output Handling	-	This vulnerability occurs when an AI output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.
Insecure Plugin Design	-	AI plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

[13] Federated learning - Wikipedia

[14] AI Glossary: Artificial intelligence, in so many words | Science

[15] Generative artificial intelligence - Wikipedia

Term	Abbreviation	Brief explanation
Language Model for Dialogue Applications	LaMDA	is a family of conversational large language models developed by Google.[16]
Large Language Models	LLM	Are where an algorithm has been trained on a large amount of text-based data, typically scraped from the open internet, and so covers web pages and - depending on the LLM - other sources such as scientific research, books, or social media posts.[17] Examples include ChatGPT, Google Bard and Meta's LLaMA.
Large Language Model Meta AI	LLaMA	A state-of-the-art foundational large language model designed to help researchers advance their work in this subfield of AI. Smaller, more performant models such as LLaMA enable others in the research community who don't have access to large amounts of infrastructure to study these models, further democratizing access in this important, fast-changing field.[18]
LLM Hallucinations	-	LLMs are also prone to "hallucinating," which means that they can generate text that is factually incorrect or nonsensical. Such hallucinations happen because LLMs are trained on data that is often incomplete or contradictory. As a result, they may learn to associate certain words or phrases with certain concepts, even if those associations are not accurate or are unintentionally "overly accurate" (by this I mean they can make up things that are true but not meant to be shared). This can lead to LLMs generating text that is factually incorrect, inadvertently overly indulgent, or simply nonsensical.[19]
Low-Level Design	LLD	Low Level Design (LLD) is specifying the HLD and describes the actual logic for the entire components of the solution. Detailed Network Security functional diagrams with all the relations and methods among all logic come under the Low-level design. Technical specifications are included with references to the HLD. LLD explains all the functional parts of the solution.

[16] LaMDA - Wikipedia

[17] ChatGPT and LLMs: what's the risk - NCSC.GOV.UK

[18] Introducing LLaMA: A foundational, 65-billion-parameter language model (meta.com)

[19] Understanding LLM Hallucinations

Term	Abbreviation	Brief explanation
Machine Learning	ML	Refers to algorithms that leverage new data to improve their ability to make predictions or decisions. ML is a widely used form of AI that has contributed to innovations such as speech recognition and fraud detection.
Machine Learning Operations	ML Ops	Is a paradigm that aims to deploy and maintain machine learning models in production reliably and efficiently. The word is a compound of "machine learning" and the continuous development practice of DevOps in the software field.[20]
Machine Learning Security Operations	ML SecOps	By leveraging artificial intelligence (AI) and machine learning (ML), security events can be identified quickly without generating low-value alerts that require analyst time, attention and manual remediation. AI and ML can identify important security events in an organization, with high fidelity, by stitching together data from multiple sources while reducing the time and experience required in the SOC.[21]
Model Denial of Service	-	Attackers cause resource-heavy operations on AI solutions, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of AI solutions and unpredictability of user inputs.
Model Theft	-	This involves unauthorized access, copying, or exfiltration of proprietary AI (LLM) models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.
Natural Language Processing	-	A computer's attempt to “understand” spoken or written language. It must parse vocabulary, grammar, and intent, and allow for variation in language use. The process often involves machine learning.[22]

[20] MLOps - Wikipedia

[21] Security Operations (SecOps) - Palo Alto Networks

[22] AI Glossary: Artificial intelligence, in so many words | Science

Term	Abbreviation	Brief explanation
Neural Network	-	A highly abstracted and simplified model of the human brain used in machine learning. A set of units receives pieces of an input (pixels in a photo, say), performs simple computations on them, and passes them on to the next layer of units. The final layer represents the answer.[23]
Node Package Manager	NPM	Is the world's largest software registry. Open-source developers from every continent use npm to share and borrow packages, and many organisations use npm to manage private development as well.[24]
OpenAI	-	Is an American artificial intelligence (AI) research laboratory consisting of the non-profit OpenAI, Inc. and its for-profit subsidiary corporation OpenAI, L.P. OpenAI conducts research on artificial intelligence with the declared intention of developing "safe and beneficial" artificial general intelligence, which it defines as "highly autonomous systems that outperform humans at most economically valuable work"[25]
Overreliance	-	Systems or people overly depending on AI solutions without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by AI solutions.
Pathways Language Model	PaLM	A Large Language Model developed by Google. It is currently in its second iteration of PaLM 2.
Prompt Injection	-	This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

[23] AI Glossary: Artificial intelligence, in so many words | Science

[24] About npm | npm Docs (npmjs.com)

[25] OpenAI - Wikipedia

Term	Abbreviation	Brief explanation
Secure Web Gateway	SWG	Also known as a web security gateway, is a device, cloud service, or application that is deployed at the boundaries of a network to monitor and stop malicious traffic from entering the organization, and to block users from accessing malicious or suspicious web resources.
Sensitive Information Disclosure	-	AI solutions may inadvertently reveal confidential data in its responses, leading to unauthorised data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.
Software Bill of Materials	SBOM	Is a nested inventory, a list of ingredients that make up software components.[26]
Supervised Learning	-	A type of machine learning in which the algorithm compares its outputs with the correct outputs during training. In unsupervised learning, the algorithm merely looks for patterns in a set of data.[27]
Supply Chain Vulnerabilities	-	LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins can add vulnerabilities.
Training Data Poisoning	-	This occurs when AI training data is tampered with, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behaviour. Sources include Common Crawl, WebText, OpenWebText, & books.

[26] Software Bill of Materials (SBOM) | CISA

[27] AI Glossary: Artificial intelligence, in so many words | Science