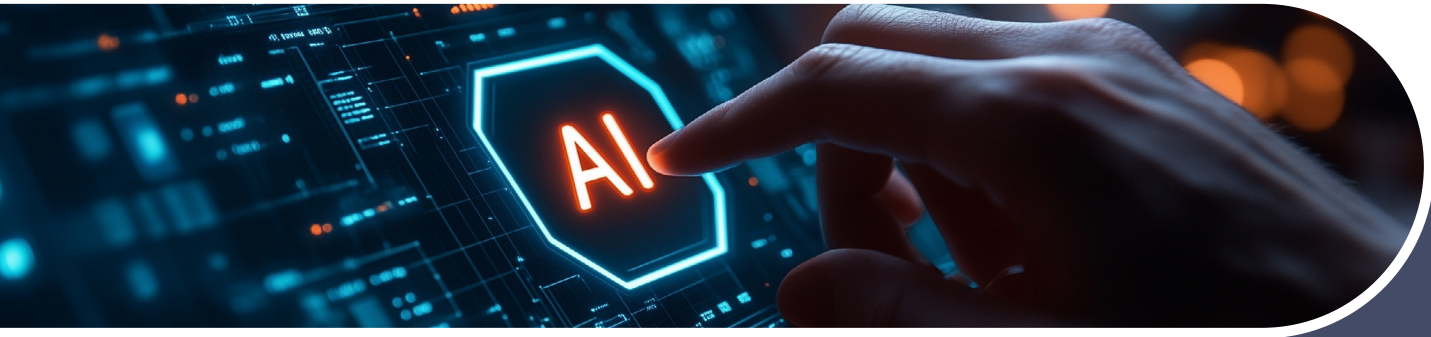
Cyberfort

WHITE PAPER

Discover how your organisation can make sure their AI Development is responsible and aligned to Secure by Design principles

Introduction



In November 2023 the UK, US and several other country governments came together to release a statement and a set of guidelines relating to AI Development and Secure by Design principles (1). The guidelines advise how those who are using AI should handle their cybersecurity when developing or using AI models as the governments who created the guidelines claimed, "security can often be a secondary consideration in this fast-paced industry."

At a high level the guidelines focused on 4 key areas of Secure by Design principles:

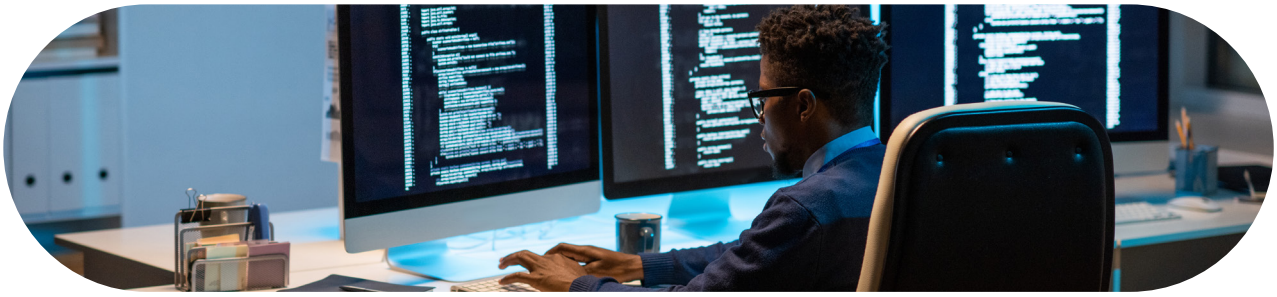
- **Secure Design:** Emphasizes designing AI systems with security in mind from the outset.
- **Secure Development:** Provides guidance for secure coding practices during system development.
- **Secure Deployment:** Addresses protecting infrastructure and models from compromise during deployment.
- **Secure Operation and Maintenance:** Covers actions relevant after deployment, including logging, monitoring, and updates.

In this whitepaper Cyberfort explores the best practices all organisations should be implementing as part of AI Development and Secure by Design principles. It identifies the key challenges organisations will face with AI, offers advice on how to start and what should be focused on from a cyber security perspective.

Understanding why AI security is different to the rest of your technology landscape

AI requires a different approach to cyber security compared to traditional cyber security methods employed. AI by its nature is dynamic as it learns from its mistakes, updates in real time, is available to everyone who has a device linked to the internet and relies heavily on data being processed by Large Language Models (LLM's) to be effective.

On the surface AI promises many benefits in terms of more productive employees, better customer service as answers to questions can be automated and delivered through a bot and enables developers to create code faster than any human for example. But there are several security risks which need to be considered when AI tools are being designed, developed, deployed and managed.



The main cyber security risks IT teams need to be aware of when reviewing AI for their organisation include:

Data security: As mentioned earlier in this paper AI tools and systems are heavily reliant on data sets. Data if not secured correctly can be vulnerable to corruption, breaches and other targeted attacks. To mitigate this risk organisations, need to make sure they have the right data security governance, policies and procedures in place to protect data integrity, confidentiality and availability throughout an entire AI lifecycle.

AI models: Attackers have been known to and are targeting publicly available and in house (through social engineering and phishing) AI models for theft, reverse engineering and manipulating data which in turn inhibits performance and reliability of AI. If an attacker is compromising an AI models integrity by interfering with its architecture, weights or parameters it can result in the AI tools and what it is being used for becoming redundant.

Input manipulation attacks: These involve altering input data to influence the behaviour and/or outcomes of AI systems. Attackers may attempt to manipulate input data to evade detection, bypass security measures or try to influence decision-making processes, which can lead to biased or inaccurate results. E.g. Data poisoning LLM's.

Adversarial attacks: Involve manipulating input data to deceive AI systems, which can lead to incorrect predictions or classifications in terms of the data used to inform an AI model. An attacker may generate different adversarial examples which exploit AI algorithms which in turn interfere with an AI model resulting in biases happening or they may send prompt injections to trick AI tools into taking harmful actions such as data leaks or deleting important company documents.

Supply chain attacks: Supply chain attacks happen when attackers target specific AI systems at the supply chain level at their development, deployment or maintenance stages. Attackers will often look for and exploit vulnerabilities in third-party components, software libraries or modules used in AI development, leading to data breaches or unauthorised access.

AI models deterioration: AI models if not regularly reviewed, updated and checked against compliance requirements can experience deterioration over time. If AI models are simply built and left unchecked this can lead to performance issues. Attackers will often look at when AI models were last updated and look for weaknesses which haven't been identified. IT security teams need to have the ability to monitor AI models for changes in performance, behaviour or accuracy to maintain their reliability and relevance.

Other security risks outside of immediate attacks on an organisation which IT teams need to be aware of with AI include:

Ethical and safe deployment: If Cyber Security teams are not prioritising safety, trust and ethics when deploying AI systems, they risk data privacy violations, biases in the information AI is producing and false positives. Ethical and trust considerations need to be in place so an organisation can ensure fairness, transparency and accountability in AI decision-making.

Regulatory compliance: As AI is so heavily reliant on data it is imperative that regulatory compliance and the management of data is considered a security risk with AI. Organisations designing, deploying and managing AI tools must comply with regulations such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA) and the EU AI Act or leave themselves open to regulatory, financial and reputational risk.

Why Secure by Design is crucial for AI system development

AI system development by its nature is complex and security needs to move from being a reactive consideration to one which is embedded across an AI systems lifecycle. In the NCSC guidelines they identify 5 key areas of focus:

- Raise staff awareness of threats and risks in the design phase.
- Model the threats to your system.
- Design your system for security as well as functionality and performance.
- Consider security benefits and trade-offs when selecting your AI model.
- Establish a security-centric culture and accountability system.

But identifying the key areas of focus in one thing. Making sure your organisation has the right skills, approach and understanding of how to implement a Secure by Design approach should not be underestimated. If Secure by Design principles are not considered as part of AI System Development organisations could be storing up potential security issues later down the line when users, developers and managers of AI systems are engaging with an AI system later on.

In the next section of this paper Cyberfort experts outline the key considerations from the NCSC guidance and offers practical advice on how to start implementing Secure By Design principles throughout an AI system development lifecycle.

01 Secure Design: Emphasizes designing AI systems with security in mind from the outset

Designing AI systems with security in mind from the outset is a crucial foundational pillar. If the design stage for an AI system is done correctly at the beginning of the strategy process it will help protect an organisation from many of the AI attacks identified in the previous section of the paper. The AI Development and Secure by Design announcement set out a number of key guidance points. But how do you take these guidance points to make sure they are being implemented and adhered to in your organisation?

In this section we review each of the guidance points for Secure Design and provide practical advice on where to get started and the key outcomes you should be looking to achieve.

When developing an AI system at the design stage it all starts with raising staff awareness of threats and risks. A conventional cyber security approach to raising staff awareness of the risks associated with AI may not be fit for purpose when it comes to informing end users, developers and managers of the risks. As a starting point those responsible for AI system development should firstly examine the business case for designing an AI tool, choice of AI tool, in-house data it will be using, type(s) of LLM selected and understand how the AI system will be used in day-to-day work.

By conducting this research, potential security threats can be identified and then guidance can be created for different user groups. The guidance should not just make staff aware of the risks but should also outline what the processes and procedures should be in case of a cyber attack on the AI system. For example, an end user in finance may just need to know at a headline level what the risks are and best practices for using an AI tool in their work. For developers it may mean having to understand which secure coding techniques they need to use and what training/certifications they need.

The next stage is to start to model threats to the AI system. This means undertaking a risk audit to understand the different threats and potential impact on an AI system. Not all cyber risk audits for AI are the same and many organisations do not have the right independent skills to complete an audit of this nature. It is recommended an NCSC assured auditor is selected for the work as they will be able to give an unbiased view of the AI system being developed and the security risks associated with AI system development. Additionally, by having an independent risk audit undertaken your organisation will benefit from wider industry knowledge and be able to take lessons learned from an NCSC auditor who will have seen many of the risks identified previously. They will also be able to assess your security maturity levels and be able to provide a roadmap for improvement aligned to your organisation's cyber security model and its maturity.

After the audit stage has taken place those responsible for AI system development can start to look at designing the AI system for functionality/performance vs security concerns based on the audit results. For example, functionality of the AI system, user experience, deployment environment, ethical, legal, and overall security posture can be assessed vs the benefits the AI tool will bring to the organisation. Trade off's in terms of risk vs performance can be identified and look to mitigated before the AI system is actually developed, third party code which may be used can be sandboxed and tested so it doesn't affect the wider organisation's security position and a range of design options based on the research can be created and scored against the risk audit. Additionally, by taking the time to review the AI system design in-life management factors can be reviewed in terms of architecture design, configuration, the different types of LLM data the AI model will use, then management, training and security parameters can be put in place before the AI system is actually developed.

02 Secure Development: Provides guidance for secure coding practices during system development

In this section of the paper, we review the Secure Development considerations of an AI system. If AI tools are being developed in the organisation without the appropriate design considerations outlined in the previous section many of the security risks, performance of the AI system and user experience will not be contemplated as part of a successful development phase. Which in turn could result in AI systems being developed which do not offer the right security protections for the organisation, deliver a great end user experience or will not have considered the training and awareness requirements as part of development. So, what are the areas your organisation should be focused on when it comes to the Secure Development stage?

It is highly likely your organisation will be using the basis of their AI system on a 3rd party set of LLM's already in existence. This means the supplier of an AI tool needs to be reviewed if the code, data and prompts you are going to use from the AI tool selected is going to be part of your overall AI ecosystem. Questions which need to be asked before using a third-party AI tool as part of your organisations AI system development include:

- Is the supplier aware of security risks associated with their AI tools?
- Where is the LLM data being sourced from, and does it conform to relevant regulatory standards?
- Does the third party you are sourcing LLM's and AI tools from follow the same/similar standards to security as your own organisation?
- Where will the risk be in terms of deploying 3rd party code into your organisation data?
- Does the supplier have well documented hardware and software components which make up their AI tool so they can be properly assessed in terms of risk, user experience and what to do if something goes wrong?
- What is the failover position of the 3rd party supplier if an attack happens? What support will they give you? What skills do they have available?
- Are they a good supplier in terms of updating the wider AI community of potential risks?
- Do they adopt NCSC guidelines as part of their AI practices or are they doing it in isolation?

By asking these questions of your 3rd party supplier for AI, your organisation can quickly assess if they will be suitable and the right fit for developing an AI system which can be secure and compliant in your operating environment.

Once the third-party supply chain risks have been identified and logged it is back to reviewing your own organisations development capabilities and assets. This includes reviewing and understanding the assets related to AI system development. Software, data, logs, documentation, and end user assessments all need to be undertaken to understand not only how they will be used in terms of developing an AI system but also where unsafe practices may occur.

At this stage of development all assets involved in AI system development need to be reviewed against confidentially, integrity and availability. Processes and controls to manage what data AI systems can access, and content generated by AI according to its sensitivity can then be created and managed. This will improve the overall security posture as all assets at this development stage will be understood and plans put in place to secure/mitigate risk against each internal asset which could be compromised by an attack on the AI system.

Once all the assets and protective processes are put in place, data, models and prompts need to be identified as part of the overall AI system. They need to be reviewed in terms of how each of type of data, models and prompts will be used throughout the AI system lifecycle. Care needs to be taken as to where data is going to be stored, used and transited as part of an AI system. Then best practices in terms of documenting training for different groups, security processes, and operations can be created as part of the AI system development.

Finally, an often-overlooked area of AI system development is the technical debt aspect. Technical debt management in terms of AI can be complex. The reason it is complex is due to the rapid development cycles to keep AI systems up to date and currently a lack of established protocols to manage. AI system architecture, code and underlying platforms will need to be reviewed and updated in almost real time to keep technical debt to a minimum or the AI tools deployed risk becoming ‘legacy’ before they can even be properly used. As part of any AI system development a technical debt process needs to be able to:

Identify and document not only where the AI system will have technical debt but also take into the account the underlying architecture, code and platforms which may be releasing new features or versions to make the AI system run effectively.

Have a planning process in place which can identify complex features/releases/updates across the AI system architecture and underlying platforms. The IT team has capacity available to prioritise technical debt in relation to an AI system, and decide on what needs refactored, retired or simply just better managed.

Have DevOps capabilities available to implement continuous integration, continuous testing, and continuous delivery of new features across all underlying platforms to make the AI system secure and productive. Review where incremental improvements can made so problems are not stored up before they become a serious legacy problem.

Make sure a structured approach to address patching and upgrades is in place. Fixing changes in code, testing and updating systems related to AI is crucial for any AI tools success. Planning the resources needed to update in a programmatic/agile/incremental way is probably the right way for most organisations. It will reduce the burden on the IT team by having to undertake all updates in one go and mitigate the impact on end users if their system needs to be taken down for a period of time for example.

Communicate and report the reasons why updates have taken place. The importance of the updates and why technical debt needs to be correctly managed in relation to AI to the wider organisation.

03 Secure Deployment: Addresses protecting infrastructure and models from compromise during deployment.

Once an AI tool has been developed and is ready to be deployed several Secure by Design principles need to be addressed. If your organisation is not in a reactive mode to AI deployment then AI systems should be released in a responsible manner only after models, applications or systems have been subjected to appropriate and effective security evaluation such as benchmarking and red teaming.

However, as AI has become more prevalent across organisations many IT teams have seen through 'Shadow IT' AI tools deployed into the organisation without the correct design or development steps being taken. If an AI system has already been deployed without going through the previous 2 steps, then the following actions need to take place immediately.

The first step when an AI system has been deployed is to review and secure the underlying platforms and infrastructure. Security controls for API's, the AI model and data being used, stored and managed need to be protected. The right Data Loss Prevention, Network, Application and Software security controls must be in place. This will involve reviewing each technology area against the AI system and identifying where security gaps may exist and then segregating different environments based on their risk profile.

Once the key areas mentioned above have been reviewed, risks identified and potentially mitigated, it is important to focus on how to protect the AI model continuously. AI as mentioned in the opening section of the paper requires a different cyber security approach. To protect the AI model continuously, a proactive approach to AI model security needs to be put in place. For example, it should be recognised across the organisation that attackers can reconstruct AI models quickly and look to exploit data which is being used. This means reviewing the AI model with an 'attack mindset' and understanding the most likely areas an attacker is going to target.

Key questions to be asked include:

- How easy is it for an attacker to replicate an AI model being used in the organisation?
- Has the data being used been correctly classified and are the right security protocols in place?
- Are the underlying platforms and infrastructure up to date in terms of patches and updated feature versions?
- Have end users of the AI model been trained correctly?
- Do developers understand the risks associated with 3rd party code being used as part of an AI model they are working on?
- Are the existing cyber security controls appropriate for the AI model?
- Have you discussed the potential risks with 3rd party suppliers, and do they have the right security approach in place to close any gaps?
- What are the control access options to each area of the AI tool?
- Should the worst happen and there is a security breach can your organisations existing incident management process cope with a breach?

AI by its nature is forever evolving. Many Incident Response models are still based on responding to an attack and reviewing historical data to mitigate the risk in the future. But as AI is evolving in real time a proactive Incident Management approach needs to be put in place for AI systems. This should include creating, simulating and testing potential attacks on an AI system to review where response gaps may exist, and the potential business impact across reputation, legal and regulatory. Data should be securely backed up and be available so the AI tool once the security risk is mitigated can be up and running again quickly with minimal impact to users. Additionally, audits should be regularly undertaken to may sure policies and procedures can be updated to minimise the risk to the organisation.

Finally, any AI system deployment must focus on end user security. After all it will be the end users who will be using the AI system in their daily working practices. They are also most likely to be vulnerable to attacks as they may not be aware of the security considerations when using an AI tool for their work.

Key steps to make sure end users are aware of the risk as part of an AI tool deployment include:

Assessing each component of an AI tool and making end users aware of the security risks.

Putting in place the most secure settings for each end user group who will be using the AI setting. For example, what data will they be allowed access to? Can you stop them from manipulating sensitive data?

Are they aware of how to store and manage data correctly?

Be transparent about the security policies in relation to AI for different user groups.

Make clear the process to follow if a security incident happens with the AI tool they are using.

Capture any changes in end user behaviour. They may not be aware that they are being socially engineered or understand the ways attackers may try to access data through them as part of them using an AI tool.

04 Secure Operation and Maintenance: Covers actions relevant after deployment, including logging, monitoring, and updates.

In the final section of the paper, we discuss how AI development can be securely operated, managed and maintained. As the usage of AI has grown so have the operational management requirements. IT teams need to have a focus not only on the design, development and deployment aspects in terms of Secure by Design but also what happens 'in life'.

This starts from the first day of deployment. The right tools need to be put in place by the IT team so they have visibility of how AI is being used, the impact on underlying systems and processes and to be able to identify any drastic changes in user behaviour which could be a sign of a cyber-attack. But visibility shouldn't just be focused on the user and system behaviours. It needs to be able to review AI system inputs in terms of the data being correctly inputted and logged. For example, is data being inputted into the AI tool which is compliant with data protection regulations, are the right alerts in place for any data which is being incorrectly used and can users be closed down if they abusing their access rights regarding data and AI tool usage.

As your organisation moves forward with an AI system, all updates need to be considered via the Secure by Design approach and assess where development and deployment risks could occur through updates to an AI model. Finally, at regular points all AI systems should be reviewed when in life to collect and share lessons learned for future AI system projects. These learnings can also be applied across other AI tools being used in the organisation to develop a better security posture in relation to AI systems.

CONCLUSION

Final thoughts

The rise of AI will continue with many organisations adopting an 'AI first strategy' in the forthcoming years. The advice and guidance on offer through the National Cyber Security Centre will be a strategically important factor for all IT teams who are looking to design, develop, deploy and manage AI tools in their organisations. But taking this advice on board is only the first step. Organisations now need to implement this advice in an effective, easy to manage way, so they do not become caught out by cyber-attacks on their AI models in the future. By answering the questions raised in this paper and putting the practical advice offered into action, organisations can unlock the power of AI in a secure, resilient and compliant manner and be ready for an ever-changing digital world.

1 https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

Discover more about Cyberfort's all-encompassing Cyber Security Services

At Cyberfort we provide a range of customers with all-encompassing Cyber Security Services. We are passionate about the cyber security services we deliver for our customers which keeps their people, data, systems and technology infrastructure secure, resilient and compliant.

Our business offers National Cyber Security Centre assured Consultancy services, Identification and Protection against cyber-attacks, proactive Detection and Response to security incidents through our security operations centre and a Secure and Recover set of Cloud solutions which keeps data safely stored, managed and available 24/7/365.

Over the past 20 years we have combined our market leading accreditations, peerless cyber security expertise, strong technology partnerships, investment in our future cyber professionals and secure locations to deliver a cyber security experience for customers which enables them to achieve their business and technology goals in an ever-changing digital world.

For more information about our Cyber Security services please contact us at the details below:

+44 (0)1304 814800 | info@cyberfortgroup.com | https://cyberfortgroup.com

We look forward to working with you