Cyberfort

WHITE PAPER

Artificial Intelligence: Managing the risk effectively

6 key areas Cyber Security teams need to review

AI - 6 key areas Cyber Security teams need to review

Artificial Intelligence (AI) – one of the most discussed topics in the world of technology. Everyday people are releasing reports, thoughts and articles on what AI can potentially do and the positive impact it will have on businesses, government organisations and society both today and into the future.

At Cyberfort we believe organisations need to take a balanced view of AI and its potential. Those who are responsible for using AI and the management of it in their organisations need to realise, that yes AI can be a force for good in terms of making organisations more productive, effective and efficient. But they also need to assess and address the potential risks and cyber security challenges which come along with it.

To help IT teams and Cyber Security professionals with their thinking around AI and the potential cyber security risks, this article will summarise 6 key areas they need to be aware of and start taking action on:

- **Data Management and Privacy vulnerabilities in LLM's**
- **Governance and Compliance of AI tools**
- **Accessing the right skills and resources to assess and respond to threats**
- **Lack of a suitable framework to assess AI risks in an organisation**
- **Threat modelling playing 'catch up'**
- **Rise of 'AI weaponry' for cyber attacks**

Understanding the current and potential cyber security risks from AI

Most people reading this article will probably understand AI and some of the potential risks associated with it. A comprehensive guide to all the associated cyber security risks related to AI doesn't currently exist according to the report released on the 15th May 2024 by the UK Department for Science, Innovation and Technology (1).

However, from our own experiences at Cyberfort over the past 12 months we have seen customers having to identify and respond to a number of common cyber security challenges in relation to AI.

01 Data Management and Privacy vulnerabilities in LLM's

As all AI projects rely on data and the quality of data provided, now is the time for organisations to review their data infrastructure and architecture. Cyber Security teams need to be working with their Information Management counterparts to regularly review how data is stored, managed and used in transit in relation to LLM's and how they are powering their AI programmes of work.

Many employees in organisations using AI tools are accessing, transforming, and annotating customer and confidential data to inform their Large Language Models (LLM's). Quite often as LLM's are difficult to build, organisations will take a baseline LLM created by a 3rd party. This means they cannot control/review the security in the model and are focused on controlling context and access.

In October 2023 OWASP published it's top 10 list of critical vulnerabilities for LLM's (2). It discovered many IT teams were not aware of how vulnerable they were in terms of attackers exploiting their LLM's in relation to data poisoning, sensitive information disclosure and supply chain vulnerabilities through 3rd party datasets.

If attackers can find ways into LLM's it obviously goes without saying that the potential risk to an organisation could be catastrophic. Cyber security teams need to review the LLM's they have in place for the AI models in their organisations and take the time to identify where vulnerabilities may exist, how they can mitigate the risk and bring them back under control. This is not an easy task given the volume of data and time involved.

This raises several questions in relation to data management and privacy such as:

- Where was the data to inform the LLM collected, managed and stored?
- Is the 3rd party providing the data for the LLM trusted? And do they comply with your own organisations data security policies and industry regulations?
- Has the data once received been reviewed correctly against Data Protection guidelines?
- If an organisation is supplementing the LLM with their data will this data be made available to a wider audience?
- How will the data in the LLM be kept safe and secure?
- Is the data used and supplemented being correctly documented at each stage?

If an organisation is not considering the above questions and adhering to the right data privacy and management guidelines in their respective industries, they could be leaving themselves open to potential data breaches, customer/employee complaints and fines from regulators.

Additionally, cyber security teams should be collaborating across the wider IT organisation to ensure data which is being used in LLM's for AI purposes is being stored correctly in both their own datacentres and in cloud computing environments. To overcome the potential data management and privacy challenges the following actions need to be undertaken when reviewing data privacy and management in LLM's:

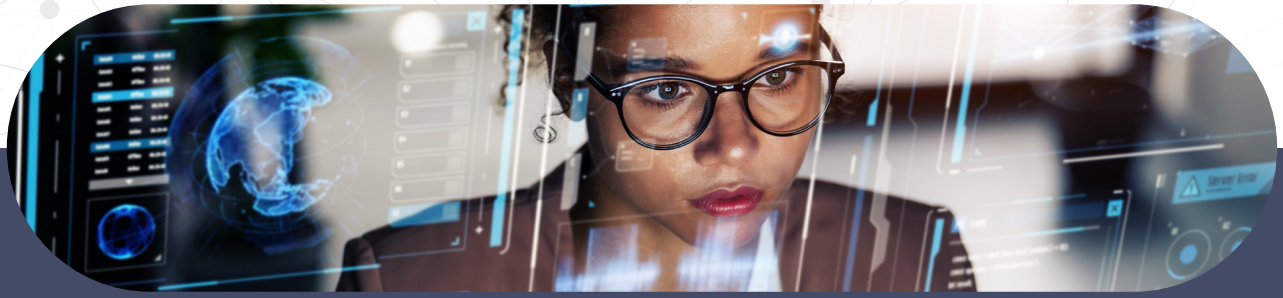
Review which LLM models are being deployed in your organisation to inform AI tools – Public API access, Licenced model, Pre-trained Model, Fine-tuned Model or Custom Model.

Assess each type of LLM model and where the data was sourced vs your own organisations data protection policies.

Review where the data security and privacy vulnerabilities may be in the LLM chosen. For example, review where data poisoning could have occurred, prompt injections may have been inserted, supply chain vulnerabilities in terms of the 3rd party supplier not adhering to certain security/regulatory policies.

If you are inputting data into an LLM assess how it will be stored, managed and who it will be made available to in the future.

Put in place security controls around each LLM such as data anonymisation, encryption of data if it is taken out of an LLM, validation of queries to prevent jailbreaking, user access controls and make sure networks are secured and cyber security monitoring tools are giving the right visibility to changes in user/machine behaviour.



02 Governance and compliance of AI tools

Many employees are using AI tools without their organisations knowledge. According to Fishbowl in 2023 (3) 70% of workers are using ChatGPT tools without their bosses knowledge. This means not only do their bosses not know what they are using ChatGPT for, but the Cyber Security team is also not aware and does not have visibility of any potential risks.

The growth of AI usage is well documented as having put pressure on many organisations IT infrastructure in terms of daily performance and technical debt. But a major issue which is being overlooked by many IT teams is how visible are end users' usage of AI tools in their organisation and how can they manage the AI tools effectively from a governance/compliance perspective?

By not understanding or having full visibility of how employees are using AI tools for their work organisations are leaving themselves open to compliance risks in highly regulated industries such as Financial Services and intellectual property risks in Manufacturing for example.

To try and be on top of this situation cyber security teams should:

Develop a view of the AI tools being used in their organisation and map to relevant compliance and regulatory frameworks. Start by identifying the different types of data, code, prompts and infrastructure being used to process information for AI purposes.

Then create and enable a systematic AI mapping capability which can review different components, measure against compliance criteria and assess the risks involved. Make sure the information from the AI audits and compliance mapping is available to senior leaders for inclusion in their risk frameworks and for formal reporting.

Once the basic visibility and mapping for compliance and governance is in place for AI create the business case for ongoing AI compliance enablement across the organisation.

Make sure enforcement of compliance and governance measures are positive by promoting trust and transparency amongst users. Open channels for 2-way feedback so people feel part of improving the security posture around AI and can proactively contribute. As AI is so vast and widespread the IT team will be heavily reliant on end users reporting issues and problems before they become a major challenge.

Become involved with the wider AI and Cyber Security ecosystem by discussing and raising threats to industry peers, take key learning's from other organisations who are using the same/similar tools as your organisation and suggest how things can be improved for a safer AI environment.

03 Accessing the right skills and resources to assess and respond to threats

It has been well recognised and reported over recent years that there is a severe Cyber Security skills shortage. AI will from what Cyberfort has seen, continue to exacerbate this trend. As the volume of data being used and created goes up, so does the risk of security vulnerabilities being exploited by attackers. But the Cyber Security industry has been slow to shift its focus to AI.

According to KPMG (4) at end of 2023 only 6% of organisations had a dedicated cyber security team reviewing AI threats and potential risks. This means most organisations are still focusing their cyber security teams and budgets on the traditional cyber security threat landscape.

To mitigate this risk IT teams need to:

Understand the different business cases for AI in the organisation and the potential resourcing pressures for the IT team.

Recognise the risks associated with different AI systems at each stage of design, development, deployment and in life management.

Review existing cyber security teams, budgets and processes and decide where they need specialist help to supplement their skills gap in relation to AI and the risks associated.

Once areas for specialist cyber security help for AI have been identified decide which skills gap approach to take - outsource, out hire or outsmart and then implement. The reality is it will be a mix of all 3 approaches to resolve the skills gap.

Understand the talent available inside and outside your organisation in relation AI security. It is likely to not be just ‘in house’ or can be rectified through training. Remote and global talent pools should be assessed in relation to each AI tool/system deployed.

Partner with industry and academic organisations to access their latest thinking and approaches as they may already have the answers you are looking for.

Look for ways to automate many of the high volume/simple cyber security tasks associated with AI. This will free up the existing IT Teams time to concentrate on more complex/strategic tasks in relation to AI management.

04 Lack of a suitable framework to assess AI risks in an organisation

In relation to the 'skills gap' problem with security and AI, many organisations are in a 'reactive mode' trying to mitigate the effects of cyber-attacks through AI tools after they have happened. Many cyber security teams are just dealing with the day-to-day issues in front of them. They are not proactively putting in place frameworks to assess their vulnerabilities in relation to AI and benchmarking their performance.
To try and help combat this reactive stance NIST produced its Artificial Intelligence Risk Management Framework in January 2023 (5). It details a best practice framework for dealing with AI across risk tolerance, management, prioritisation and integration of AI risk management across an organisation.

At Cyberfort we recommend those who are responsible for reviewing AI risks evaluate the suggested NIST framework, identify gaps and benchmark their maturity levels. To start developing your own AI risk framework 5 key steps should be undertaken:

Create and establish a capability model which has clear leadership, business objectives and AI strategy documented. Then review the AI ideation in terms of Data Discovery, Hypothesis testing, Experimentation and Creative AI. Following these steps review the actual AI system delivery on a case-by-case basis in terms of the model set up, model training, Machine Learning operations, scalability and how the AI model will be optimised and maintained when in - life. Finally, review the AI system against a trust framework which considers Ethics, Governance and Explainability before adding in how you will measure the impact of AI and how the AI system will be adopted.

Develop key risk objectives and indicators. Risk objectives need to align with the broader organisation's goals in relation to AI and the overarching risk profile. Indicators need to feed data and information into each risk objective to assess how each AI system is performing vs risk parameters. For example, risk indicators could include data quality and the percentage of datasets passing automated quality controls, model drift by measuring AI performance in terms of accuracy and precision of results, recalls against data samples or retraining by looking at the number of automated retraining cycles completed and improved model performance post training.

Data quality and LLM operations need to be reviewed at every stage. Data which is being used in LLM's needs to be of a high quality and the right governance measures put in place. Each data set informing an LLM needs the right data architecture in place so any discrepancies can be identified and rectified before the LLM becomes poisoned and does not perform how it should. An operational chain of command should be put in place from data processing to model training/updating to deployment and ethical considerations. Then each stage of the operating model should start to be reviewed against the organisations risk framework.

Gen AI can also be used as a positive tool to help the IT team to govern its own AI environments. For example, you could use your LLM's to analyse AI regulations and industry best practices from around the world and compare them with your control environment policies. The results will determine which legislation is relevant to each control, and how it is supported by your organisation.

Adopt a continuous improvement model to identifying risks in your AI systems. Review where governance and compliance requirements need to be considered and where issues are highlighted look at where they can be improved on a case-by-case basis. See this continuous improvement as a positive. Many risks when rectified result in improved user experiences and better security postures for an organisation.

05 Threat modelling playing ‘catch up’

Many cyber security professionals know how to use threat models for traditional threats and will be using a common framework such as STRIDE, MITRE and ATT&CK. However, as AI usage has exponentially grown, does your cyber security team have the time to be able to correctly create, test and manage threat models in an ever-changing AI landscape?

At Cyberfort we have witnessed many organisations struggling to keep pace with the development and management of their new threat models in relation to AI. By not having the time to create, test and develop the right threat models organisations could be missing crucial vulnerable areas which attackers could look to exploit through AI.

For example, when deploying a new AI system have you asked how your current threat model will react to the following questions being asked of it:

- If the data was poisoned in the LLM how would you know?
- If you are training your data model on 3rd party data what steps are taking to make sure security connecting your own organisations data and the publicly sourced LLM is in place?
- Have you classified the sensitivity of your data which will be used for informing an LLM and what are the security procedures if there is an attack?
- Does the current threat model understand the lineage of data?
- Does the current threat model understand and take into account the most likely types of attacks such as forcing emails to be classified as spam, attacker-crafted inputs which reduce the confidence level of correct classification, attackers injecting noise into the source data being classified, or contamination of training data to force the misclassification of select data points.
- Can your current threat model deal with the most common AI/ML attacks including - adversarial perturbation, targeted data poisoning, model inversion attacks, membership inference attacks, model stealing, neural net reprogramming, adversarial attacks in physical domains, malicious ML providers, supply chain attacks, backdoor machine learning, exploitation of software dependencies.

To develop the right AI threat model for your organisation questions like the above should be asked and then reviewed against the existing cyber security threat model. If gaps are found then the AI threat model should be reviewed against NCSC best practice (7).

06 Rise of ‘AI weaponry’ for cyber attacks

As AI becomes more sophisticated and used more frequently, it is likely many organisations will see more AI-powered attacks. According to Deep Instincts 4th Annual Voice of SecOps report 2023 (6)75% of security professionals have witnessed an increase in cyber attacks over the past year and 85% were powered by generative AI.

Attacks will range from deepfake social engineering to automated malware and hackers gaining access to detailed information about their targets. Many organisations are unprepared for these increased level of attacks in terms of the skilled people they have available and their current technology in place.

The main types of AI powered attacks include:

Scouting, scanning and analysing massive amounts of data to identify vulnerabilities in networks and systems.

Exploiting development as AI can automatically generate and adapt exploits to specific vulnerabilities, increasing the potential of attacks being successful.

Speed of attack movement as AI can move through networks more efficiently, allowing attackers to gain access to critical resources and sensitive information.

Social Engineering as AI can be used to create convincing phishing emails, social media profiles/posts which are difficult to detect and prevent.

DDoS attacks as AI can create massive botnets and coordinate their activities, launching DDoS attacks to harm websites and networks.

Recommended steps to prevent AI powered attacks include:

Making sure a layered security approach is adopted by deploying multiple security solutions at different points of the organisations network to protect against a variety of threats. This includes multi-factor authentication, encryption on data at rest and in transit, and implementation of firewalls as a minimum standard.

Use AI-powered security tools to detect and respond to AI-based cyber-attacks. These tools can help to analyse network traffic, identify abnormal behaviour, and predict potential attacks.

Implement strong authentication and authorisation controls across different user groups to prevent unauthorised access to your systems and data.

Educate different user groups on AI threats and what they should be looking for. Tailor the training to each group. For example, a Finance Administrator will need only basic training on AI security in terms of phishing, social engineering and data protection, compared to a developer who is using code in their daily work to update AI systems who will probably need to understand a range of different attacks and vulnerabilities in terms of where they are accessing and updating code.

Keep up to date on the latest AI security threats by working with technology partners, suppliers and the wider AI ecosystem to identify what threats are out there, how risks can be mitigated and contribute to the community for improved learning.

Develop an AI incident response plan which will help to quickly and effectively respond to AI powered cyber-attacks.

It is important as we have discussed in previous points that Cyber Security leaders have the right AI cyber security framework in place and can proactively plan for this unprecedented rise in attacks and the potential risks, they may be vulnerable to. This may mean in the short-term having to partner with a specialist cyber security services provider to identify and respond to the rise in attacks whilst your organisation starts to put the foundations in place for the future.

CONCLUSION

Next steps

This article has covered 6 key challenges facing cyber security teams in relation to AI. At Cyberfort we provide a range of expertise to help organisations assess their security and risk vulnerabilities in relation to AI. At Cyberfort we can work with your cyber security teams to identify the key threats of AI being deployed and used in your organisation, right through to detecting and responding to real time attacks caused by AI.

1 https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence

2 https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf

3 https://www.fishbowlapp.com/insights/70-percent-of-workers-using-chatgpt-at-work-are-not-telling-their-boss/

4 https://kpmg.com/us/en/media/news/kpmg-generative-ai-2023.html

5 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

6 https://info.deepinstinct.com/voice-of-secops-v4-2023

7 https://www.security.gov.uk/guidance/secure-by-design/activities/performing-threat-modelling

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 (01304 814800 | info@cyberfortgroup.com | https://cyberfortgroup.com

We look forward to working with you