AI and cyber security: what you need to know

Understanding the risks - and benefits - of using AI tools.

Ignited by the release of ChatGPT in late 2022, artificial intelligence (AI) has captured the world's interest and has the potential to bring many benefits to society. However, for the opportunities of AI to be fully realised, it must be developed in a safe and responsible way, especially when the pace of development is high, and the potential risks are still unknown.

As with any emerging technology, there's always concern around what this means for security. **This guidance is designed to help managers, board members and senior executives (with a non-technical background) to understand some of the risks - and benefits - of using AI tools**.

Managers don't need to be technical experts, but they should know enough about the potential risks from AI to be able to discuss issues with key staff.

What is artificial intelligence?

> **Artificial intelligence** (AI) can be described as *'Any computer system that can perform tasks usually requiring human intelligence. This could include visual perception, text generation, speech recognition or translation between languages.'*

One of the most notable recent AI developments has come in the field of **generative AI**. This involves AI tools that can produce different types of content, including text, images and video (and combinations of more than one type in the case of 'multimodal' tools). Most generative AI tools are geared towards specific tasks or domains. For example, ChatGPT effectively allows users to 'ask a question' as you would when holding a conversation with a chatbot, whereas tools such as DALL-E can create digital images from natural language descriptions.

It appears likely that future models will be capable of producing content for a broader range of situations, and both Open AI and Google report success across a range of benchmarks for their respective GPT-4 and Gemini models. Despite this broader applicability, there remains no consensus on whether artificial general intelligence – the dystopian vision of the future where an autonomous system surpasses human capabilities – will ever become a reality.

How does AI work?

Most AI tools are built using **machine learning** (ML) techniques, which is when computer systems find patterns in data (or automatically solve problems) without having to be explicitly programmed by a human. ML enables a system to 'learn' for itself about how to derive information from data, with minimal supervision from a human developer.

For example, **large language models** (LLMs) are a type of generative AI which can generate different styles of text that mimic content created by a human. To enable this, an LLM is 'trained' on a large amount of text-based data, typically scraped from the internet. Depending on the LLM, this potentially includes web pages and other open source content such as scientific research, books, and social media posts. The process of training the LLM covers such a large volume of data that it's not possible to filter all of this content, and so 'controversial' (or simply incorrect) material is likely to be included in its model.

Why the widespread interest in AI?

Since the release of ChatGPT in December 2022, we've seen products and services built with AI integrations for both internal and customer use. Organisations across all sectors report they are building integrations with LLMs into their services or businesses. This has heightened interest in other applications of AI across a wide audience.

The NCSC want **everyone** to benefit from the full potential of AI. However, for the opportunities of AI to be fully realised, it must be developed, deployed and operated in a secure and responsible way. Cyber security is a necessary precondition for the safety, resilience, privacy, fairness, efficacy and reliability of AI systems.

However, AI systems are subject to novel security vulnerabilities (described briefly below) that need to be considered *alongside* standard cyber security threats. When the pace of development is high – as is the case with AI – security can often be a secondary consideration. Security must be a core requirement, not just in the *development* phase of an AI system, but *throughout its lifecycle*.

It is therefore crucial for those responsible for the design and use of AI systems – including senior managers – to keep abreast of new developments. For this reason, the NCSC has published AI guidelines designed to help data scientists, developers, decision-makers and risk owners build AI products that function as intended, are available when needed, and work without revealing sensitive data to unauthorised parties.

What are the cyber security risks in using AI?

Generative AI (and LLMs in particular) is undoubtedly impressive in its ability to generate a huge range of convincing content in different situations. However, the content produced by these tools is only as good as the data they are trained on, and the technology contains some serious flaws, including:

- it can get things wrong and present incorrect statements as facts (a flaw known as 'AI hallucination')
- it can be biased and is often gullible when responding to leading questions
- it can be coaxed into creating toxic content and is prone to 'prompt injection attacks'
- it can be corrupted by manipulating the data used to train the model (a technique known as 'data poisoning')

Prompt injection attacks are one of the most widely reported weaknesses in LLMs. This is when an attacker creates an input designed to make the model behave in an unintended way. This could involve causing it to generate offensive content, or reveal confidential information, or trigger unintended consequences in a system that accepts unchecked input.

Data poisoning attacks occur when an attacker tampers with the data that an AI model is trained on to produce undesirable outcomes (both in terms of security and bias). As LLMs in particular are increasingly used to pass data to third-party applications and services, the risks from these attacks will grow, as we describe in the NCSC blog 'Thinking about the security of AI systems'.

How can leaders help ensure that AI is developed securely?

The Guidelines for Secure AI System Development, published by the NCSC and developed with the US's Cybersecurity and Infrastructure Security Agency (CISA) and agencies from 17 other countries, advise on the design, development, deployment and operation of AI systems. The guidelines help organisations deliver *secure outcomes*, rather than providing a static list of steps for developers to apply. By thinking about the overall security of systems containing AI components, stakeholders at all levels of an organisation can prepare to respond to system failure, and appropriately limit the impact on users and systems that rely on them.

Crucially, keeping AI systems secure is as much about *organisational* culture, process, and communication as it is about *technical* measures. Security should be integrated into all AI projects and workflows in your organisation from inception. This is known as a 'secure by design' approach, and it requires strong leadership that ensures security is a *business* priority, and not just a technical consideration.

Leaders need to understand the consequences to the organisation if the integrity, availability or confidentiality of an AI-system were to be compromised. There may be operational and reputational consequences, and your

organisation should have an appropriate response plan in place. As a manager you should also be particularly aware of AI-specific concerns around data security. You should understand whether your organisation is legally compliant and adhering to established best practice when handling data related to these systems.

It's also important to note that the burden of using AI safely should **not** fall on the individual users of the AI products; customers typically won't have the expertise to fully understand or address AI-related risks. **That is, developers of AI models and systems should take responsibility for the security outcomes of their customers**.

In addition to the AI Guidelines, the NCSC's Principles for the security of machine learning (published in 2022) provide context and structure to help organisations make educated decisions about where and how it is appropriate to use ML, and the risks this may entail. Some of the principles are particularly relevant to those in senior decision making and executive or board level roles. These are highlighted in the quick reference table on the front page of the principles.

Questions to ask about the security of your organisation's AI systems

Managers, board members and senior executives can use the following questions in discussions with technical and security staff, to help you understand how your organization is dealing with the AI/ML threat.

- Do you understand where accountability and responsibility for AI/ML security sit in your organisation?
- Does everyone involved in ML deployment, including board members and/or senior executives, know enough about AI systems to consider the risks and benefits of using them?
- Does security factor into decisions about whether to use ML products?
- How do the risks of using ML products integrate into your existing governance processes?

- What are your organisation's critical assets in terms of ML and how are they protected?
- What is the worst case (operationally or reputationally) if an ML tool your organisation uses fails?
- How would you respond to a serious security incident involving an ML tool?
- Do you understand your data, model and ML software supply chains and can you ask suppliers the right questions on their own security?
- Do you understand where your organisation may have skills or knowledge gaps related to ML security? Is a plan in place to address this?

Further reading

If you'd like to know more about AI, the NCSC has produced a series of relevant publications which are summarised below.

- Guidelines for secure AI system development: Guidelines to help developers make informed decisions about the design, development, deployment and operation of their AI systems.
- The near-term impact of AI on the cyber threat: An NCSC assessment focusing on how AI will impact the efficacy of cyber operations and the implications for the cyber threat over the next two years.
- Lindy Cameron at Singapore International Cyber Week: In this speech, the CEO of the NCSC covers the reshaping of cyber security that's required in the era of generative AI.
- ChatGPT and large language models: what's the risk?: A blog that discusses some of the cyber security considerations to be aware of when using LLMs.
- Thinking about the security of AI systems: A technical blog that looks at 'prompt injection' attacks and other vulnerabilities within LLMs.
- Exercise caution when building off LLMs: A blog aimed at cyber security professionals that explains why our understanding of LLMs is still 'in beta', and what this means for cyber security.

- Principles for the security of machine learning: Detailed guidance to anyone developing, deploying or operating a system with a machine learning (ML) component.

PUBLISHED

13 February 2024

REVIEWED

13 February 2024

VERSION

1.0

WRITTEN FOR

Cyber security professionals

Large organisations

Small & medium sized organisations

Public sector