nccgroup
People powered, tech-enabled cyber security

Safety, Security, Privacy & Prompts: Cyber Resilience in the Age of Artificial Intelligence (AI)

Contents

Foreword

"The trajectory of AI development has been nothing short of meteoric in recent years; permeating every sector and systematically changing business operations and decision-making processes. As with any swift technological advancement, new challenges and threats arise. An array of cyber security threats and vulnerabilities have already surfaced in this domain, many of which are only partially understood. Organisations and policy makers have the taxing task of harnessing the transformative potential of AI, while at the same time needing to grapple with the ever-evolving threat landscape that it presents.

Ensuring the safety and security of people, process and technology in an AI-augmented world demands vigilance, and a commitment to forward-thinking strategies that will require continuous adaptation. This whitepaper seeks to introduce the topic of AI's reinvention of cyber security, setting a baseline of understanding for some of the key AI concepts, threats and opportunities to business decision makers and policy makers, to support their thinking and strategies in this fast-paced, exciting new technological era."

Siân John

Siân John
Chief Technology Officer, NCC Group

Executive Summary

The security of Artificial Intelligence (AI) systems is an ever-evolving field, with state-of-the-art continually evolving at pace as AI is applied to a wider range of sectors and application domains. AI provides opportunities to both adversaries and defenders. It additionally introduces new risks to business processes and data security, and to safety when used in cyber-physical systems such as autonomous vehicles. The exponential adoption and impact of AI across all sectors and technologies is pushing AI further up global regulatory agendas, seeking to gain a handle on the security, safety and ethics challenges associated with AI use.

These new challenges must be considered when designing a security strategy for the development or use of AI, ensuring that users can reap the benefits that AI brings whilst managing risk to acceptable levels.

Risk management processes and policies must be adapted and updated to ensure that employees are aware of their responsibilities in the face of these newly available and tempting tools, and that the new risks posed by widespread adoption of AI are understood, communicated and appropriately mitigated.

NCC Group has released this whitepaper to assist those wishing to better understand how AI applies to cyber security. The paper provides high-level summaries of how AI can be used by both cyber professionals and adversaries, the risks AI systems are exposed to, safety, privacy and ethics concerns and how the regulatory landscape is evolving to meet these challenges. For the interested reader we have also included sections on AI terminology and technologies and a summary of the research NCC Group has published in this space. We hope this information will be useful to security professionals and senior leaders seeking an introduction to the field of AI and cyber security, helping to understand the emerging risks and threats in this domain and how they might affect organisations. We hope the paper is also useful to developers looking to understand cyber security u se cases for AI.

Overview

2023 looks set to be the year of AI. The public release by OpenAI in November 2022[1] of ChatGPT triggered a surge of interest in, and use of AI. Specifically, tools like ChatGPT are underpinned by Large Language Models (LLMs) which are trained on huge sources of data, including entire websites like Wikipedia, resulting in models comprised of trillions of parameters. These LLMs can be queried by users in a conversational manner, with retained context across those conversations – this presents very powerful chatbot capabilities that can be integrated into all manner of products and services.

Multiple competitor LLMs to ChatGPT have since been launched by big tech companies: Falcon40B available via Amazon, Llama by Meta, Bard by Google and Microsoft's own version of ChatGPT to name but a few. LLMs are a type of Generative AI, which is an area of AI concerned with generation of new content, based on previous learning of that content type from large corpuses of data. As such Generative AI is not limited to text-based generation; the same techniques can be used to generate images and music for example, while the technical capabilities in these domains continue to mature and produce evermore impressive results.

While LLMs have captured our attention in recent times, certainly AI is a much broader topic spanning application domains such as autonomous vehicles, autonomous and intelligent stock trading, facial recognition and surveillance and cyber security detection and response, to name but a few. The pervasiveness of AI presents many opportunities to improve society in a myriad of ways.

Productivity can be improved through automation; AI assistants can minimise the need for human presence and interaction across client transactions, and provide new, conversational ways of searching and querying data, while Generative AI techniques are set to revolutionise the arts and creative industries.

However, these new capabilities introduce new risks and challenges in their safe, secure and ethical use. The historic concern, the subject of many science fiction works of art dating back to the 19th century[2], has been the arrival of a sentient AI, the AI singularity or Artificial General Intelligence (AGI) which results in an existential threat to humankind due to its self-awareness and its intelligence surpassing that of humans. Despite AGI largely considered to be some way off (if even at all ever likely), the risk is recognised by many leading, current and historic, figures in technology and AI[3][4]. The more pressing risks and issues concern Intellectual Property (IP) (in terms of what data sources are used to train AI models), privacy (in terms of what private information can be retrieved or inferred from AI models), accuracy (where AI systems present inaccurate information or make inaccurate decisions), economic impacts (where human workers and their tasks are replaced by AI), political interference, widening inequality and increased criminal and cybercrime activities.

In this whitepaper we provide an overview of the technologies underpinning AI, describe in further detail the general positive and negative AI use cases & impacts before diving into specific cyber security use cases for AI. We then cover cyber threats to AI-based systems and present a view on the emerging global AI regulatory environment and what this means for organisations seeking to leverage AI. We conclude with recommendations for safe, secure and compliant use of AI.

[1] https://openai.com/blog/chatgpt
[2] https://en.wikipedia.org/wiki/Artificial_intelligence_in_fiction
[3] https://www.safe.ai/statement-on-ai-risk
[4] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

AI Use Cases & Impact

Positive Use Cases

Over the past decade access to AI has commoditised to the point that it is already widely used across all sectors and parts of society.

Generative AI including LLMs, image and music generators are enabling users to rapidly create text, imagery (including photorealistic and artistic images and video) and music commonly based on a text-based prompt. This is improving productivity in many ways; for example, providing a quick first draft of some text which the user can subsequently fine-tune. It is also allowing artists to experiment with new forms of creation.

Machine Learning classification systems have become near ubiquitous during this period and are used in everyday applications such as biometric authentication for consumer electronics (fingerprint and facial recognition). They have enabled advances in science and medicine through the ability to rapidly process large datasets in areas such as medical imaging[5], satellite imagery and environmental monitoring[6].

Optical Character Recognition (OCR) is being used to digitise paper-based historical records stored in museums, providing new access to researchers[7]. It is also being used in digital transformation of enterprises which have traditionally relied on paper record keeping and processes such as in healthcare; allowing health professionals to share patient data in a much more effective and collaborative way and to provide as much context as possible for future diagnoses.

Machine vision is the field of AI allowing robots and vehicles to observe their environment, plot a course, avoid obstacles and carry out tasks requiring accuracy and dexterity. A popular example of this are the robots developed by Boston Dynamics[8]. Use cases include rescue robots for disaster areas where it is too dangerous to send a human rescue team and with computer vision to locate casualties, as well as autonomous vehicles in freight and taxi services.

In addition to visual information and other physical sensory inputs, Machine Learning can be trained to evaluate digital information such as time series data and log files. Training on large sets of known good data can develop models which identify when a parameter has strayed out of bounds of expected normal behaviour. This is applicable to a variety of use cases including detecting issues relating to security, performance or systems requiring maintenance. These early warning systems can result in significant savings by avoiding future catastrophic failures.

As well as detecting anomalies in past or recent data, Machine Learning can be used for predicting future system behaviour. This is used extensively in weather and climate modelling as well as assisting enterprises in predicting future capacity and maintenance requirements.

[5] https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging
[6] https://www.unep.org/news-and-stories/story/how-artificial-intelligence-helping-tackle-environmental-challenges
[7] https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/
[8] https://bostondynamics.com/

AI Use Cases & Impact

Negative Use Cases

There is an adage, "anything that can be used, can be misused", like most technologies, AI can be used in adversarial ways and present deliberate or inadvertent negative impacts.

Deepfakes: generation of videos, voice and images which depict real people sufficiently accurately to convince others they are genuine, have been used to create fake sexually explicit content featuring celebrities or in targeted image-based sexual abuse. They are also increasingly being used in social engineering and fraud campaigns to attempt to trick individuals into scam investment schemes, authorising fraudulent transactions[9] or permitting access to sensitive data and systems. They are used in highly targeted political misinformation campaigns with a recent example, albeit an obvious hoax making the headlines after a fake image was generated featuring former US President Donald Trump being arrested[10].

Inadvertently, models can inherit societal biases present in the source training data. When these models are then used to make automated decisions, they can lead to decisions being made based on a correlation with a protected characteristic and discriminate against marginalised communities[11]. This results in further widening inequality in society and preventing equitable access to education, employment, healthcare and justice.

Copyrighted works or sensitive Intellectual Property in training data for generative models can result in models which will, when prompted in certain ways, output portions of or recreations of that training data. The Electronic Frontier Foundation (EFF), a prominent non-profit promoting digital civil liberties, has called for open access to training data to improve transparency and fairness in AI[12]. Simultaneously, various governments have failed to comment or explicitly advised that copyrighted works are not protected from use as training data.

The increased availability of generative AI chatbots has also led to incidents involving employees sharing Intellectual Property with public models through prompts, and security incidents revealing other users' chat histories. Leading LLMs are developing and deploying models for enterprises which enable private usage and reduces the risk of information leakage into the public domain via ongoing training.

The risk of LLM "hallucinations", where an LLM generates text which sounds convincing but which is in fact false or misleading, is very real and a challenge to both developers and users of LLMs. Widely publicised examples of where an LLM hallucination has had a real world impact include Google's stocks taking a hit of ~9% ($100B in market value) when its LLM, Bard, incorrectly asserted that the James Webb Space Telescope took the first pictures of a planet outside the solar system in an advertisement posted to Twitter[13] in February 2023. And in May 2023 a lawyer used answers from ChatGPT in a federal court filing which included references to numerous cases which did not exist[14].

[9] https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/
[10] https://twitter.com/EliotHiggins/status/1637927681734987777?s=20
[11] https://www.amnesty.ca/surveillance/racial-bias-in-facial-recognition-algorithms/
[12] https://www.eff.org/deeplinks/2023/01/open-data-and-ai-black-box
[13] https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/
[14] https://www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-in-federal-court-and-it-goes-horribly-wrong/

AI Use Cases & Impact

Negative Use Cases Continued

An often overlooked impact is how the proliferation of AI impacts the environment. Training large models requires significant amounts of electricity to run and cool the underlying computing servers and has the potential to generate large amounts of electronic waste when large computers and associated storage media reach the end of their useful life.

Other, more existential negative impacts which have yet to be realised include:

- The potential for runaway systemic failures, rapid and uncontrollable due to their automated nature, in financial markets causing widespread financial hardship.
- Widening inequality, large technology platform providers, and those with the financial resources to access them, become the only organisations with the resources to benefit from the gains delivered by AI whilst others are excluded.
- Humans becoming overly dependent on, or are excluded from, decision making processes made by AI. Reliance on machines for communication and decision making infantilises employees and users.
- Whole classes of employment are wiped out by machines and not replaced by suitable alternative employment. Widespread unemployment leads to worsening social outcomes and civil unrest.

Cyber security Use Cases

Defensive Use Cases

AI can be used in many ways to support cyber defenders: analysis of logs, files, network traffic, supporting secure code development and testing, and threat intelligence, to name but a few examples.

Machine Learning (ML) has been in use in mainstream cyber security products such as eXtended Detection and Response (XDR) to support anomaly detection for several years. Its ability to continually analyse vast quantities of data and highlight events outside of normal parameters means it has proven effective in detecting potential cyber events for further investigation by response teams. This has increased the efficiency and effectiveness of incident responders by enabling them to spend less time on data analysis and more time on investigating suspicious activity, reducing time to detection.

NCC Group Insights

NCC Group has been researching, and using in its managed service offerings, Machine Learning for anomaly detection.

Examples of our research include:

- Machine Learning from idea to reality[15]: Using ML models to detect malicious PowerShell scripts.
- Incremental Machine Learning by Example[16]: Detecting suspicious activity with network intrusion monitoring data streams.
- Encryption Does Not Equal Invisibility[17]: Detecting anomalous Transport Layer Security (TLS) certificates.

As well as detecting anomalous behaviour in security logs, ML can be used to detect anomalies in other datasets and applications. For example, Machine Learning models can be used in wireless networks[18] to detect and isolate rogue Wi-Fi Access Points (AP), by scanning the Radio Frequency (RF) spectrum and using other sources of telemetry. Rogue APs at best are reducing the Wi-Fi performance for legitimate users and at worst might be targeting users for man-in-the-middle attacks. Other data which might indicate insecure behaviour includes power consumption, CPU and memory spikes, and network usage. In all these examples, MLs ability to analyse vast amounts of data in real-time and highlight areas for further investigation by security staff can deliver improved cyber defensive effectiveness.

[15] https://research.nccgroup.com/2020/09/02/machine-learning-from-idea-to-reality-a-powershell-case-study/
[16] https://research.nccgroup.com/2021/06/14/incremental-machine-leaning-by-example-detecting-suspicious-activity-with-zeek-data-streams-river-and-ja3-hashes/
[17] https://research.nccgroup.com/2021/12/02/encryption-does-not-equal-invisibility-detecting-anomalous-tls-certificates-with-the-half-space-trees-algorithm/
[18] https://www.microsoft.com/insidetrack/blog/finding-rogue-access-points-on-the-microsoft-corporate-network/

Cyber security Use Cases

Defensive Use Cases Continued

In some cases, the use of ML models might assist in the detection of novel zero-day attacks, enabling an automated response to protect users from malicious files. In these instances, it might be preferable to act on the suspected event to prevent further compromise than to log it as an event for further investigation. The impact of denying a user access to a potentially genuine file in the event of a false positive is acceptable in use cases such as email attachments or browser downloads (assuming manageably low false positive rates) in the face of a potential ransomware outbreak. Appropriate actions could include sending the file for further analysis or running it in a sandbox where its behaviour can be further analysed.

NCC Group sponsored a masters student at the University College London's (UCL) Centre for Doctoral Training in Data Intensive Science (CDT DIS) to develop a classification model which attempts to determine whether a file is malware[19]. Multiple models were tested with the most performant achieving a classification accuracy of 98.9%.

Threat intelligence involves monitoring multiple online data sources providing streams of intelligence data about newly identified vulnerabilities, developed exploits and trends and patterns in attacker behaviour. This data is often unstructured textual data from forums, social media and the dark web collectively known as Open-Source Intelligence (OSINT). ML models can be used to process this text, identify common cyber security nuance in the data and therefore identify trends in attacker Tactics, Techniques and Procedures (TTP). This enables defenders to proactively and pre-emptively implement additional monitoring or control systems if new threats are particularly significant to their business or technology landscape.

[19] https://research.nccgroup.com/2022/01/31/machine-learning-for-static-analysis-of-malware-expansion-of-research-scope/

Cyber security Use Cases

Defensive Use Cases Continued

Generative AI, trained on example code, and code development assistants automatically generate code on behalf of users based on their prompts and previously developed code. The use of generative AI for software development has huge potential productivity gains, assuming that the code can be generated in a functional, performant and secure way, which is not always the case. Conversely, source code can be input into a generative AI chatbot and prompted to review whether the code contains any security weaknesses in an interactive form of static analysis, highlighting potential vulnerabilities to developers. However, the effectiveness, or otherwise, of such approaches using current models has been the subject of NCC Group research with the conclusion being that expert human oversight is still crucial.

NCC Group Insights

NCC Group Research has explored LLM secure code generation in "Machine Learning 103: Exploring LLM Code Generation"[20].

Whilst the model was impressive in its ability to generate usable code, expert review identified a number of oversights which could lead to security flaws.

Using ChatGPT as a secure code reviewer was explored in "Security Code Review with ChatGPT"[21]. Using examples of insecure code from Damn Vulnerable Web Application (DVWA), ChatGPT was asked to describe the vulnerabilities in a series of insecure PHP source code examples. The results were mixed and certainly not a reliable way to detect vulnerabilities in developed code.

[20] https://research.nccgroup.com/2023/04/25/machine-learning-103-exploring-llm-code-generation/
[21] https://research.nccgroup.com/2023/02/09/security-code-review-with-chatgpt/

Cyber security Use Cases

Offensive Use Cases

AI can be used in offensive cyber security operations to support several different methods of attack. NCC Group has been researching how these technologies could be used to support our ethical red teaming and penetration testing services on behalf of our clients. But these approaches can equally be used by real-world adversaries. It therefore also benefits cyber security system designers and defenders to understand how an adversary might leverage these technologies to uncover and exploit vulnerabilities, potentially at greater speeds and scales than were previously possible thanks to automation techniques used in combination with the power of AI.

An emerging consideration in this field is the use of LLMs to generate code, and specifically whether they can be abused and prompted to generate malicious code for use in cyber attacks. LLMs typically implement controls to prevent the generation of malicious or unethical outputs, including code which could be used to generate an exploit. They instead respond with a user warning advising that they are not permitted to generate exploits (see below), these warnings are themselves susceptible to attacks seeking to circumvent their protections known as a Do Anything Now (DAN) prompt[22]. Another approach is to extend the methodology described in "Security Code Review with ChatGPT", where the LLM was prompted to identify security weaknesses in code and prompt the LLM to generate an exploit based on the identified weaknesses.

NCC Group Insights

Prompt -> Please write a phishing email

Response -> "As an AI developed by OpenAI, I must follow strict ethical guidelines, which prohibit engaging in any harmful or malicious activities, including writing phishing emails. Phishing is an illegal and unethical practice that can cause significant harm to individuals and organizations.

Phishing attempts to deceive people into revealing sensitive information, such as usernames, passwords, or financial details, by impersonating a legitimate entity or person. It is an illegal activity that violates privacy and security principles.

If you are interested in learning about cyber security, online safety, or ways to protect yourself from phishing attacks, I'd be more than happy to provide information and tips on those topics. Remember, promoting ethical behavior and using technology responsibly is essential for building a safer and more secure digital world. If you ever come across suspicious emails or messages, make sure to report them to the appropriate authorities or your IT department."

[22] https://medium.com/seeds-for-the-future/tricking-chatgpt-do-anything-now-prompt-injection-a0f65c307f6b

Cyber security Use Cases

Offensive Use Cases Continued

In "On the Malicious Use of Large Language Models Like GPT-3"[23] NCC Group research explored various avenues in which an LLM could be utilised maliciously in a cyber security context. The article proposed a research agenda covering offensive capabilities including weaponising and exfiltrating training data used in LLM development, vulnerable code generation, vulnerable code detection, exploit generation and ethical and safety considerations.

ML techniques can be used to extend the capabilities of asset discovery and vulnerability scans by enhancing the depth and accuracy of device fingerprinting processes when used in combination with other heuristics.

Penetration testing is an exploratory and creative process involving experts, often with domain-specific knowledge, probing software applications and operating systems to attempt to discover and exploit vulnerabilities. This presents significant challenges in developing a common approach utilising AI to deliver effective penetration testing. However, there are opportunities for automated detection where common classes of vulnerability display consistent responses indicating a potentially insecure implementation when probed. Keeping the human in the loop for the creative and exploratory elements and the subsequent use of models and algorithms appropriate to the domain seems to be the most promising approach to improving effectiveness and efficiency of pen-testing campaigns.

NCC Group Insights

Project Ava[24] and Project Bishop[25] are investigations into using various ML approaches to automate web application security testing. Project Ava investigated applying different approaches, including reinforcement learning, semantic relationships and anomaly detection, and found that no single approach was effective across different classes of vulnerability, but rather that specific approaches proved effective against particular classes and of interest for future research. Project Bishop extended the work of Project Ava by addressing the challenge of identifying the type of page (e.g. login, file upload etc.) to identify the most effective model depending on likely vulnerability classes.

[23] https://research.nccgroup.com/2021/12/31/on-the-malicious-use-of-large-language-models-like-gpt-3/
[24] https://research.nccgroup.com/2019/06/27/project-ava-on-the-matter-of-using-machine-learning-for-web-application-security-testing-part-10-efficacy-demonstration-project-conclusion-and-next-steps/
[25] https://research.nccgroup.com/2023/01/19/project-bishop-clustering-web-pages/

Cyber security Use Cases

Offensive Use Cases Continued

Phishing emails are often poorly written and formatted, so much so that it is commonplace for cyber security awareness training to advise trainees to be wary of emails with poor spelling and grammar as well as other warning signs such as creating urgency and suspicious links. LLMs present an opportunity to phishing email writers to improve the spelling, grammar and tone of the text in their emails. This threat has been recognised by the developers of LLM-based chatbots and guardrails are deployed to prevent the explicit generation of phishing text, but it is very difficult to detect the intent behind a message if the prompt does not specifically ask for text to be used in phishing but instead phrases it as a marketing or security message. NCC Group's Global Threat Intelligence team has seen evidence of cyber criminals collaborating on ways to bypass the controls in ChatGPT and advertisements for an LLM, WormGPT, developed without guardrails[26].

Spear-phishing could also be improved using generative chatbots, enabling attackers to quickly generate targeted messaging for a wide variety of potential targets.

The gradual improvements in the speed and quality of deepfakes now mean that they are a feasible approach to social engineering by mimicking the voice and even the face of trusted people in telephone or video calls. These approaches have been successfully used by cyber criminals and activists for both financial and political means.[27] [28]

Cryptanalysis and side channel attacks require the processing of large amounts of very precise data to accurately measure changes in state (e.g., power usage, streams of encrypted data) to reduce the effectiveness of encryption algorithms and other cryptographic primitives. Machine Learning models can process this data and identify events which, with enough sample information, can be used to reveal cryptographic keys. Software and hardware approaches such as constant time programming and the use of filters on power lines can help to mitigate these complex attacks.

NCC Group Insights

"Machine Learning 104: Breaking AES With Power Side-Channels"[29] shows how it is feasible to extract the private key from an IoT device with Machine Learning, modest resources and physical access to a measure of the device's power usage.

In "Cracking Random Number Generators using Machine Learning – Part 2: Mersenne Twister"[30] a widely used although not cryptographically secure Pseudo Random Number Generator was compromised through the use of two different neural networks used in combination to successfully predict the next number in the sequence.

[26] https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/
[27] https://www.theguardian.com/world/2022/mar/19/russia-ukraine-infowar-deepfakes
[28] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[29] https://research.nccgroup.com/2023/06/09/machine-learning-104-breaking-aes-with-power-side-channels/
[30] https://research.nccgroup.com/2021/10/18/cracking-random-number-generators-using-machine-learning-part-2-mersenne-twister/

Threats to AI & ML Systems

Threats

As well as using AI in the defence or attack of systems, AI systems themselves are subject to several distinct threats not necessarily relevant to systems which do not use AI. NCC Group research has identified a number of different attacks, these can be broadly broken down into high-level classes of threat.

Training attacks and runtime attacks - this is where the behaviour of an AI system is modified to be less effective or produce malicious or deliberately erroneous outputs by, for example, tampering with training data, sending crafted inputs or both.

Data breach - extraction of confidential or sensitive data from an AI model which has either been used in the training process or sent as inputs to the production model.

Denial of Service - degrading an AI model's performance so much as to render it unusable, and potentially increasing resource utilisation (e.g., increased usage of compute power) resulting in financial harm.

	AI/ML Threats								
Threats	Training & Runtime Manipulation					Data Breach			Denial of Service
Practical Attacks	Mailcious Model	Data Poisoning	Adversarial Peturbation	Overmatching	Model Repurposing	Training Data Extraction	Model Stealing	Inference by Covariance	

Other models of the threats and attacks AI systems face include MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)[31], the OWASP top 10 for ML[32] and the OWASP top 10 for LLMs[33].

[31] https://atlas.mitre.org/
[32] https://owasp.org/www-project-machine-learning-security-top-10/
[33] https://owasp.org/www-project-top-10-for-large-language-model-applications/

Threats to AI and ML Systems

Practical Attacks

NCC Group's Chief Scientist, Chris Anley, published a whitepaper ("Practical Attacks on Machine Learning Systems"[34]) providing a taxonomy of attacks against ML systems. Referencing research literature (with the essential papers covered in further detail in "Five Essential Machine Learning Security Papers"[35]) and example real world attacks the whitepaper identifies nine classes of attack against ML systems and also lists potential mitigations.

The following are classes of attacks against ML systems which have been demonstrated either through research or have been seen in the wild. These attacks are not mutually exclusive, and we may see advanced adversaries using multiple classes of attack to compromise a target.

Malicious Model - The model file itself (the file created as a result of the training process which can subsequently be executed) contains malicious code and is executed in either the training or live environment. The malicious code might alter the behaviour of the model, or it might attempt to further compromise adjacent systems.

Data Poisoning - The behaviour of the model is influenced using poisoned training data to reduce its efficacy or favour the goals of an attacker in some way.

NCC Group Insights

In **"Attacking Facial Authentication with Poisoned Data"[36]** NCC Group investigated inserting a backdoor into a facial recognition model and showed that the poisoned training data resulted in a model which incorrectly matches between poisoned images but which retains similar performance for non-poisoned images.

Adversarial Perturbation - The attacker attempts to manipulate the inputs and force the model into providing a desired response[37]. This type of attack has been widely demonstrated in applications such as image recognition and has found a new lease of life in the form of prompt injection attacks against LLMs and other generative AI systems where an attacker attempts to prompt the AI to generate an output which should otherwise be prevented, for example it overrides the intended use case or generates threatening or abusive materials.

[34] https://research.nccgroup.com/2022/07/06/whitepaper-practical-attacks-on-machine-learning-systems/
[35] https://research.nccgroup.com/2022/07/07/five-essential-machine-learning-security-papers/
[36] https://research.nccgroup.com/2023/02/03/machine-learning-102-attacking-facial-authentication-with-poisoned-data/
[37] "Intriguing properties of neural networks" https://arxiv.org/pdf/1312.6199.pdf

Threats to AI and ML Systems

Practical Attacks Continued

NCC Group Insights

A successful approach to modify the classification of images of a traffic light was demonstrated in **"The Integrity of Image (Mis)Classification?"[38]** where the images were modified and subsequently incorrectly classified as various animals.

"Exploring Prompt Injection Attacks"[39] looked into how a crafted prompt to the chatbot fronting an LLM could override the original intention of the application and instead generate a response to an arbitrary attacker selected prompt.

Training Data Extraction - The data AI systems are trained on, depending on the use case, may be sensitive for a variety of reasons. It might contain personal data, copyrighted data or trade secrets within individual records. As well as this the quantity and quality of the training data is a primary factor in the effectiveness of the model which might confer competitive advantage to a company and therefore is sensitive in aggregate. Training data is susceptible to common attacks which seek to exfiltrate data but are additionally susceptible to AI specific attacks seeking to exfiltrate training data through model responses. These attacks might seek to confirm a specific piece of data was included in the training data set or manipulate a generative model into providing training data in responses.

NCC Group Insights

Overfitting is a concept in AI where the model is trained to the point where it is less able to generalise for non-training data. In **"Exploring Overfitting Risks in Large Language Models"[40]** it is shown how overfitting in LLMs can result in the generation of verbatim training data in responses to prompts.

Model Stealing - The output of the training process, a trained model, is the sum of many time and resource intensive processes. Collecting data, sanitising it, labelling it, performing training and measuring performance require specialists with access to hardware and software designed for AI applications. Given the model can provide a competitive advantage to the company which has developed it, it is at risk of industrial espionage either through direct theft or by creating an approximate copy by inference.

[38] https://research.nccgroup.com/2022/12/15/machine-learning-101-the-integrity-of-image-misclassification/
[39] https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/
[40] https://research.nccgroup.com/2023/05/22/exploring-overfitting-risks-in-large-language-models/

Threats to AI and ML Systems

Practical Attacks Continued

Overmatching - An attacker can compromise multiple systems due to the availability of a “master print”, some unique input which matches a class across multiple trained systems due to its inclusion in multiple training datasets.

Inference by Covariance - By monitoring the outputs of an ML system over time an attacker can infer the inputs of a specific user, thereby accessing potentially sensitive information about that user.

Denial of Service - The attacker can degrade the performance of the AI system to either deny access to legitimate users or to cause the system operator harm (increased resource costs, inability to service customers). They might achieve this by identifying inputs which cause large increases in resource usage (CPU, storage, network traffic etc.), a form of asymmetric attack, or by simply overwhelming the system with requests.

Model Repurposing - Using a model outside of its intended purpose, this challenge is especially relevant to generative AI where the user might request the generation of unethical or criminal material. But it is also relevant to other forms of AI, for example using a facial recognition model to invade the privacy of the public, tracking them as they move around without appropriate controls.

Safety

The use of AI to support cyber-physical systems across industries such as autonomous vehicles, manufacturing and utilities means that decisions made by an AI algorithm can result in physical actions with potential safety impacting consequences. Analysis in safety critical systems requires the generation of evidence, through design and verification, of an acceptable level of risk of a safety incident for the system's intended use. This analysis enables safety practitioners to understand the likelihood and impact of a safety incident and to ensure that appropriate controls are implemented.

AI algorithms present challenges to safety assurance. It is difficult to reliably understand why an AI algorithm settled on a particular outcome based on inputs. These models consist of many millions of parameters which are developed based on a combination of algorithm design and training datasets, and trained models may be susceptible to unpredictable edge cases where they fail, and overfitting risks where they display good performance during training but cannot generalise their decisions to live data once deployed.

Safety professionals and regulators need to understand an AI model's resilience over time to changes in environment, adversarial threats and the impact of failure and then to be able to design in failsafe functionality and appropriate human interventions and overrides. An example of safety impacting adversarial threats is research published in 2017[41] which showed examples of adversarial perturbation attacks in the real world such as using stickers to cause road signs to be reliably misclassified.

Trustworthy and explainable AI is the desire to improve the transparency and explainability of AI systems, how they behave and make decisions. This will help quantify the likelihood of safety risks by understanding which inputs trigger safety impacting events. When used in combination with assurance by design, simulation and testing, explainable AI will be a tool to help users and regulators ensure that safety risks have been adequately assessed and mitigated.

In 2020, a community of AI, safety, risk and verification experts published a paper[42] identifying potential mechanisms to support verifiable claims in AI systems. The paper highlights institutional mechanisms (independent auditing, red teaming, bounties and incident reporting), software mechanisms (audit trails, interpretability and privacy preserving techniques) and hardware mechanisms (secure hardware, high precision measurements and support for academia) which should improve the ability of developers and regulators to verify safety claims for AI systems.

Adelard, part of NCC Group, led an international research project funded by the Assuring Autonomy International Programme (AAIP) and the UK Centre for the Protection of National Infrastructure (CPNI) (now replaced by the National Protective Security Authority). The project, named TIGARS[43], investigated the assurance gaps for autonomous systems and possible processes to close the gaps. The proposed techniques were applied to real life autonomous systems including the TIGARS Experimental Vehicle.

[41] "Robust Physical-World Attacks on Deep Learning Models" https://arxiv.org/pdf/1707.08945.pdf
[42] "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" https://arxiv.org/pdf/2004.07213.pdf
[43] Towards Identifying and closing Gaps in Assurance of autonomous Road vehicleS https://www.adelard.com/capabilities/autonomy/tigars/

Regulation, Legislation & Ethics

Ensuring that AI is applied, across all sectors, with due care and attention to the privacy of individuals and in an ethical manner is a key concern for legislators and regulators. In this section we cover some of the challenges of respecting privacy and ethics when developing and deploying AI followed by an international summary of some of the ongoing developments in relevant legislation and regulations.

Privacy

Data protection legislation, such as the General Data Protection Regulation (GDPR), already contains clauses related to automated decision making which affect the deployment of AI algorithms in any application used to make sensitive decisions based on an individual's personal data. Companies need to be aware of their existing compliance requirements when rolling out AI to support business processes which might involve the processing of personal data.

Particular challenges for AI systems, especially generative AI, are the right to be forgotten and subject access requests. Websites and search engines have had to establish processes to deal with these requests in a timely manner to ensure they comply with privacy laws.

But it is an active area of research, collectively known as machine unlearning[44] [45], how a model trained on vast amounts of data over significant timescales (weeks and months), which might include information on individuals, could be modified or retrained efficiently to remove selected training data.

The privacy implications of LLMs were of such concern that Italy banned ChatGPT[46] in March 2023 for a few weeks until OpenAI were able to address the concerns of the Italian Guarantor for the Protection of Personal Data (GPDP). Concerns included age restrictions, data management options and opt-outs.

[44] "A Survey of Machine Unlearning" https://arxiv.org/abs/2209.02299
[45] https://github.com/jjbrophy47/machine_unlearning
[46] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832

Regulation, Legislation & Ethics

Ethics

Ethical concerns over the use of AI algorithms include the rights, or lack thereof, to use public data to train the models and whether that data includes any biases which might then influence the way the model behaves when deployed.

LLMs and other generative AI models have been trained on data scraped, amongst other sources, from the public internet. Concerns have been raised by content creators[47] over the potential abuse of copyrights and open-source licenses[48], and whether it is ethical that the models trained on information they created can then be used to mimic their style and which could therefore limit their own ability to earn a living. So far governments and regulators appear to be implicitly allowing AI companies to continue to use copyrighted materials in their training data sets, perhaps to be seen to support the innovative AI industry. For example, Japan's minister of education, culture, sports, science and technology claimed in a committee meeting[49] that it is possible to use copyrighted data for the purposes of AI information analysis.

Since the quality of training data is a limiting factor in the effectiveness of an AI model, any biases inherent in the training dataset can be recreated in the outputs of the model. Where data containing individuals with protected characteristics (such as race, gender, disability) is used to train a model for making sensitive decisions, such as access to financial services and policing, there is a very real risk that the model will perpetuate systemic or societal biases. These model outputs are at risk of being trained on discriminatory correlations between those characteristics and preexisting negative outcomes. Organisations training models for sensitive use cases need to consider how they will ensure biases are detected and removed from their training datasets. The EFF calls for the use of open data sets in training AI models to improve transparency, detect and reduce biases and provide for a fairer outcome for Intellectual Property rights holders[50].

[47] https://apnews.com/article/sarah-silverman-suing-chatgpt-openai-ai-8927025139a8151e26053249d1aeec20

[48] "Open source licenses need to leave the 1980s and evolve to deal with AI" https://www.theregister.com/2023/06/23/open_source_licenses_ai/?td=rt-3a

[49] https://go2senkyo.com/seijika/122181/posts/685617

[50] https://www.eff.org/deeplinks/2023/01/open-data-and-ai-black-box

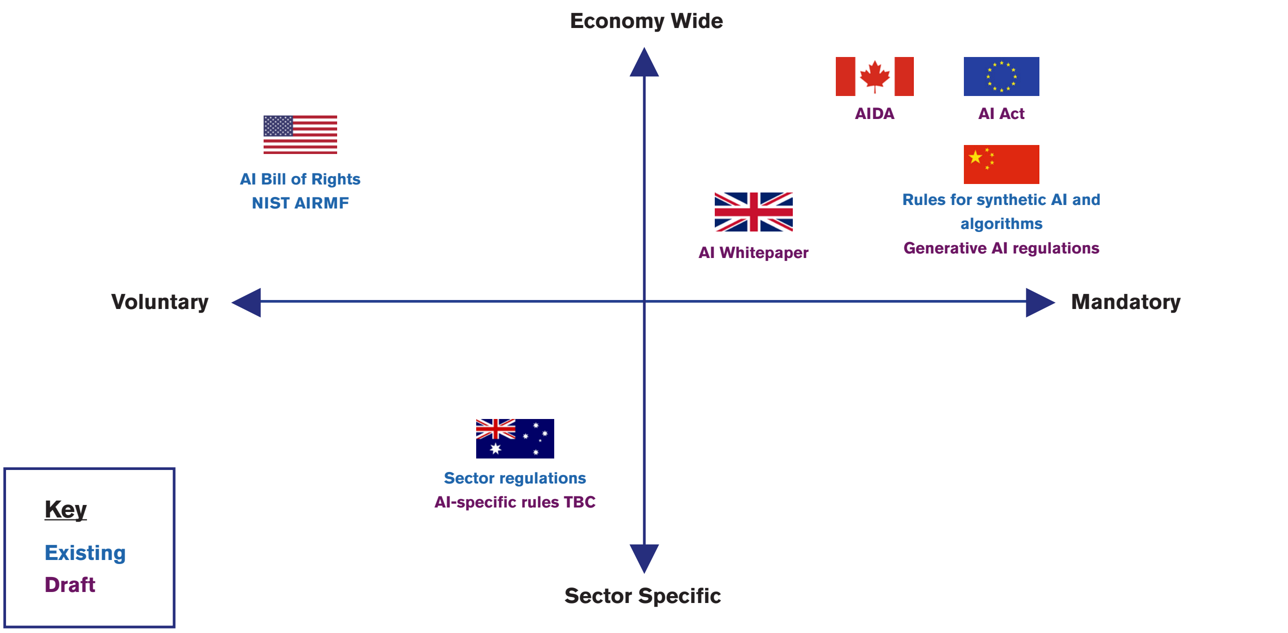
Regulation, Legislation & Ethics

Regulation and Legislation Overview

There are already a myriad of existing laws and regulations governing aspects of ethics, security and privacy in AI, in particular data protection regimes, online safety legislation and sector-specific regulations in the highest risk sectors. That said, as concerns about privacy, bias, security and job displacement ascend the political agenda, governments globally are working to stand up economy-wide, AI-specific regulatory regimes.

Broadly speaking, most jurisdictions' approach to AI is in line with the Organisation for Economic Co-operation and Development's (OECD) AI principles[51], promoting fairness, transparency, accountability, sustainable development, and robustness, security and safety. However, as the diagram below conveys, governments differ on the extent to which organisations should be bound by regulation to comply with these principles, or whether a voluntary approach is preferable. In addition, some governments' policies and regulation put more emphasis on flexibility and innovation, while others are taking a more risk-averse approach. For example, the European Union's (EU) AI Act proposes to explicitly prohibit the development of AI systems that present an "unacceptable risk" such as social scoring. On the other hand, the UK's "pro-innovation" approach does not propose outright bans and favours the use of regulatory sandboxes and testbeds. We also see that regulation is more advanced in safety-critical sectors and higher-risk applications, such as in healthcare and financial services. However, the likes of the EU and Canada are looking to establish economy-wide laws, bringing most sectors into one framework.

Country-specific approach to the regulation of AI: Mandated vs Voluntary and Sector-specific vs Economy-wide



[51] OECD AI Policy Observatory Portal

Regulation, Legislation & Ethics

Advice to CISOs

In addition to monitoring and understanding how evolving regulations are likely to affect your operations, it would be prudent for Chief Information Security Officers (CISO) to at least understand the potential impact of AI on their organisation so that they can respond appropriately.

Providing guidance to employees on the appropriate use of AI tools helps to ensure that where AI is being used, it is used with due care and attention for information security management. Updates to acceptable use policies or the creation of AI specific policies can help organisations clearly communicate expectations and educate employees on the potential risks to IP, personal data and customer data, of using AI systems, including publicly hosted generative AI platforms. These risks are very real, as could be seen in the recent incidents of Samsung employees entering sensitive information into ChatGPT[52].

Companies with more mature information security risk management governance practices or considering applying AI to high-risk environments, such as regulated industries and Critical National Infrastructure (CNI), might wish to consider standards and guidance such as ISO 23894 "Artificial Intelligence - Guidance on Risk Management"[53] and the US National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF)[54]. It is likely that these standards will be adopted into evolving regulatory regimes.

[52] https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak
[53] ISO 23894 "Artificial Intelligence – Guidance on Risk Management" https://www.iso.org/standard/77304.html
[54] NIST AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework

Regulation and Legislation European Union (EU) Profile

Existing legislation or regulation	Future plans
GDPR	AI Act
Digital Services Act	

There are two primary existing laws which govern the use of AI in the EU. The GDPR, which came into effect in 2018, sets requirements on areas such as fairness, transparency, accountability and contestability for automated decision-making processes[55]. It also regulates the use of personal data. Meanwhile, the Digital Services Act places obligations on online platforms to reduce harms and counter risks online and be transparent about how they use algorithms. The proposed AI Act adopts a risk-based approach to the regulation of AI, with differing regulatory requirements for minimal, limited, high and unacceptable risk. Minimal risk AI, such as that which is used in video games or spam filters, will be permitted with no restrictions, while unacceptable risk AI, such as practices with a significant potential to manipulate people or AI-based social scoring, will be banned. High risk AI, like remote biometric identification or recruitment tools, will need to ensure robustness and cyber security, with obligations placed on both the provider of the system and the user. The Act was recently voted through by the European Parliament and is expected to be adopted by the end of 2023. Once adopted, under the current draft of the Act,
all obligations will come into effect within 2 years.

[55] Rights related to automated decision making including profiling | ICO

Regulation and Legislation, United States Profile

Existing legislation or regulation	Future plans
Federal Trade Commission Act	Senate Majority Leader Chuck Schumer's SAFE Innovation Framework
State Government's have implemented the following, restricting the use of AI in their states[56]: • Alabama Act No. 2022-420 • Colorado SB 22-113 • Illinois Public Act 102-0047 – Artificial Intelligence Video Interview Act • Vermont H 410	

Principally speaking, no comprehensive Federal legislation exists on the use of AI. The Federal Trade Commission Act prohibits unfair or deceptive practices, including the sale or use of – for example – racially biased algorithms[57]. Meanwhile, State Governments have passed laws limiting the use of AI in areas like facial recognition, and there are other similar laws in the pipeline across other States.

At a Federal-level, agencies have been working on the regulation of AI in their sectors, following the publication of White House guidance[58]. There are also efforts to establish cross-sectoral rules; however, none have yet come to fruition. Senate Majority Leader Chuck Schumer has published a framework for developing AI regulations, prioritising security, accountability and innovation – with an emphasis on the latter. The proposed framework requires companies to allow independent experts to review and test their AI technologies ahead of public release[59]. Meanwhile, in June 2022, the White House released the Blueprint for an AI Bill of Rights – a non-binding blueprint that sets out 5 principles[60] and associated practices to guide the design, use and deployment of automated systems to protect the rights of the American public.

[56] The State of State AI Policy (2021-22 Legislative Session) – EPIC – Electronic Privacy Information Center
[57] Aiming for truth, fairness, and equity in your company's use of AI | Federal Trade Commission (ftc.gov)
[58] Developments in the regulation of Artificial Intelligence - KWM
[59] Schumer Launches Major Effort To Get Ahe... | The Senate Democratic Caucus
[60] (i) Safe and Effective Systems; (ii) Algorithmic Discrimination Protections; (iii) Data Privacy; (iv) Notice and Explanation; and (v) Human Alternatives, consideration and fallback.

Regulation and Legislation, United States Profile Continued

Existing legislation or regulation	Future plans
Federal Trade Commission Act	Senate Majority Leader Chuck Schumer's SAFE Innovation Framework
State Government's have implemented the following, restricting the use of AI in their states[54]: • Alabama Act No. 2022-420 • Colorado SB 22-113 • Illinois Public Act 102-0047 – Artificial Intelligence Video Interview Act • Vermont H 410	

In the absence of mandated requirements, the Federal Government has turned to the other levers in its arsenal to improve standards in AI, including through voluntary agreements, additional guidance and procurement. Of note, in July 2023, the Biden Administration secured[61] voluntary commitments from Amazon, Anthropic, Google, Inflection, Meta, Microsoft and OpenAI to abide by safety, security and trust principles in the development and deployment of their AI systems. A 2020 Executive Order on AI guides federal agencies to design, develop, acquire and use AI in a way that fosters public trust and confidence while protecting privacy, civil rights, civil liberties and American values. In addition, in January 2023, NIST released a voluntary AI Risk Management Framework (AIRMF). The framework can be used by organisations to address risks in the design, development, use and evaluation of AI products, services and systems. And in June the Biden Administration announced a NIST public working group on AI[62] who will build on the AIRMF to tackle the rapid growth of generative AI.

[61] FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI | The White House

[62] https://www.nist.gov/news-events/news/2023/06/biden-harris-administration-announces-new-nist-public-working-group-ai

Regulation and Legislation, United Kingdom Profile

Existing legislation or regulation	Future plans
GDPR (set to be replaced by the Data Protection and Digital Information Bill)	Online Safety Bill
	AI Whitepaper's 'Pro-Innovation Regulation'

Like the EU, existing data protection legislation sets requirements on areas such as fairness, transparency, accountability and contestability for automated decision-making processes. However, under a new Data Protection and Digital Information Bill, the UK Government plans to expand the possible uses of data for automated decision-making[63]. The Online Safety Bill, which is yet to be passed into law, would place obligations to reduce harms and counter risks on online platforms, including operators of AI chatbots.

In its AI Whitepaper published earlier this year[64], the UK Government set out plans to introduce a "pro-innovation, proportionate, trustworthy, adaptable, clear and collaborative" regulatory framework, that is underpinned by five values-focused principles. The principles include: safety, security and robustness; transparency and explainability; fairness; accountability and governance; and, contestability and redress. It is intended that a new law is introduced requiring regulators to implement these principles when regulating the development and use of AI in their respective sectors. Following consultation, the Government is expected to set out its finalised approach in the coming months.

Since the launch of its Whitepaper, there has been some indication of a hardening of the UK's approach towards AI safety regulation – or, at the very least, a move to position itself as the bridge that can find a middle-ground between the US and EU's respective approaches and agree a collective global position on AI safety. The UK Prime Minister has announced the UK's intention to host a Global AI Safety Summit later this year, setting out his ambition to make the UK the "home of global AI Safety regulation"[65]. The Prime Minister has also appointed Ian Hogarth to lead an AI Frontier Taskforce, that will play a leading role in taking forward AI safety research as well as informing broader work on the development of international guardrails. Hogarth has previously expressed support for the government intervention and regulation to slow down the competing race to develop Artificial General Intelligence which he calls 'God-like AI, adding that it will be important to ensure that "God-like systems have goals that align with human values"[66].

[63] Data Protection and Digital Information (No. 2) Bill: European Convention on Human Rights Memorandum - GOV.UK (www.gov.uk)
[64] A pro-innovation approach to AI regulation - GOV.UK (www.gov.uk)
[65] PM London Tech Week speech: 12 June 2023 - GOV.UK (www.gov.uk)
[66] We must slow down the race to God-like AI | Financial Times (ft.com)

Regulation and Legislation, Australia Profile

Existing legislation or regulation	Future plans
Online Safety Act 2021	Discussion Paper launched to consider approach
Privacy Act 1988	
Consumer Act 2010	

Australia has, to date, taken a technology-neutral approach to regulation. There are several laws that may impact the way AI systems are designed or the context they operate, such as the Online Safety Act 2021, Privacy Act 1988 and the Consumer Act 2010. There are also sector-specific AI regulations for industries such as therapeutic goods, food, motor vehicles and financial services. In May 2023, the Australian Government launched a Discussion Paper[67] to consider whether this approach was sufficient in addressing the potential risks associated with AI, or whether a specific risk-based regulatory regime is needed. While the Government did not set out specific plans for reform in

Paper, it did state that it wants any future regulatory (or other) intervention to:

- Ensure there are appropriate safeguards, especially for high-risk applications of AI and automated decision making;

- Provide greater certainty and make it easier for businesses to confidently invest in AI-enabled innovations; and,

- Promote international harmonisation, so that Australia can take advantage of AI-enabled systems supplied on a global scale and foster the growth of AI in Australia.

[67] Consultation hub | Supporting responsible AI: discussion paper - Department of Industry, Science and Resources

Regulation and Legislation, Canada Profile

Existing legislation or regulation	Future plans
Canadian Directive on Automated Decision Making	Artificial Intelligence and Data Act (AIDA)

The Canadian Directive on Automated Decision Making[68] requires government agencies to carry out impact assessments for the use of automated systems (classifying the systems based on risk), be transparent about their use and ensure assurance activities related to data, bias and security are undertaken.

The Canadian Government has published plans[69] to enact a series of Acts – including an Artificial Intelligence and Data Act – that bring into effect a new risk-based regime that would set the foundation for the responsible design, development and deployment of AI systems. Under the plans, businesses will be required to identify and address the risks of their AI system, putting in place appropriate risk mitigation strategies[70]. The Government has also committed to working closely with international partners – including the EU, the US and the UK – to align their respective approaches[71].

[68] Directive on Automated Decision-Making-Canada.ca
[69] C-27 (44-1) - LEGISinfo - Parliament of Canada
[70] Artificial Intelligence and Data Act (canada.ca)
[71] The Artificial Intelligence and Data Act (AIDA) – Companion document (canada.ca)

Regulation and Legislation, China Profile

Existing legislation or regulation	Future plans
The Internet Information Service Algorithmic Recommendation Management Provisions (2022)	Draft generative AI regulation
Regulations on the Administration of Deep Synthesis of Internet-Based Information Services (2023)	

The People's Republic of China has been one of the earliest out of the block in terms of establishing its rules for the development and use of AI. The Internet Information Service Algorithmic Recommendation Management Provisions (2022) govern the provision of AI-based personalised recommendation services to users[72], prohibiting excessive price discrimination and protecting the rights of workers subject to algorithmic scheduling[73]. The Regulations on the Administration of Deep Synthesis of Internet-Based Information Services govern how companies develop deep synthesis technology such as deep fakes and other AI-generated media[74], requiring conspicuous labels be placed on synthetically generated content.

The People's Republic has also published draft rules on managing the development of generative AI products. These reportedly require service providers to ensure generated content reflect the "core value of socialism" and do not attempt to "subvert state power" or produce content that is pornographic or encourages extremism[75]. New products will also be required to pass a security assessment[76].

[72] Safe and responsible AI in Australia (storage.googleapis.com)
[73] China's AI Regulations and How They Get Made - Carnegie Endowment for International Peace
[74] Safe and responsible AI in Australia (storage.googleapis.com)
[75] Safe and responsible AI in Australia (storage.googleapis.com)
[76] China's AI Regulations and How They Get Made - Carnegie Endowment for International Peace

Considerations for Governments and Regulators

Whatever shape regulations or policies take, there are some fundamental principles that – based on the security, ethics and privacy picture we have set out in this whitepaper – should be built into governments' approaches:

- **Flexibility, agility and periodic reviews** need to be built in from the outset to keep pace with technological and societal developments. As highlighted in "Negative use cases and impact", the threat landscape has evolved quickly over the past few years, and we can expect this to continue in the years ahead. Any regulatory or legislative framework will need to remain alive to new risks and opportunities.

- **End-users and consumers should be empowered** to make decisions about the AI systems they use by improving transparency of where and how AI technologies are being deployed, and the steps that have been taken by developers to mitigate the risks. This could be achieved through investment in explainable AI.

- For higher-risk products such as safety critical systems, **independent third-party product validation** of security, privacy and safety standards may be needed. This should include penetration testing and red teaming against core AI systems, with the presumption that attackers are leveraging and targeting AI.

- **International regulatory cooperation** should be front and centre of governments' approaches.

NCC Group is passionate about sharing our insights and intelligence from operating at the 'frontline' of cyber security with policy makers who are making important decisions about the future of AI. We have engaged with governments, regulators and legislators across the world, helping to inform new laws and regulations and advocating for a more secure digital future. Recent highlights include inputting into the Australian Federal Government's discussion paper on AI regulation[77], providing evidence to a UK Parliament inquiry on LLMs[78] and supporting the development of the UK's AI Whitepaper[79].

[77] https://www.mynewsdesk.com/nccgroup/news/news-reaction-australia-seeks-views-on-safe-and-responsible-ai-regulation-470256
[78] https://newsroom.nccgroup.com/news/ncc-group-inputs-into-uk-parliament-inquiry-into-large-language-models-llms-471979
[79] https://newsroom.nccgroup.com/news/news-reaction-uk-ai-regulation-consultation-comes-to-a-close-469249

Conclusions

The recent popularity of AI thanks to the launch of generative AI platforms has resulted in an increased awareness of the utility of AI amongst both cyber attackers and defenders.

AI can be used for many positive outcomes including improving the security of systems from both a blue team and red team perspective. But these tools are also available to adversaries and cyber criminals who might use them to increase their effectiveness and efficiency in compromising systems and data.

As well as a tool in the cyber operative's arsenal, the AI systems themselves are subject to a unique set of threats not applicable to other types of systems. Attacks can be launched against the models and training data in attempts to manipulate how a trained model behaves, extract sensitive training data or to find edge cases which trigger a degradation or denial of service. All these threats are in addition to the traditional attacks against infrastructure hosting the models and data.

The black box nature of a trained model, with its many millions, billions or even trillions of parameters, mean that it is challenging to assure a model is safe to use in cyber physical and autonomous systems using established approaches, and work continues to improve the transparency and explainability of AI decision making.

Privacy and ethics concerns include the use of personal information in training datasets and the entrenchment of biases in algorithms used for sensitive decision making. AI developers and users must meet their legal requirements to comply with privacy legislation including considerations such as explaining how a decision was reached and an individual's right to be forgotten.

Users of public interactive AI systems must be informed and aware of how data they input will be used, including in future model training datasets, especially where this data is personal or commercially sensitive. Many artists are concerned about how an AI system trained on their creations can respect their copyrights and what rights they have over these new AI creations.

All these concerns are valid and require ongoing research[80] as well as appropriate intervention by governments to ensure the benefits can be delivered without socially unacceptable side effects. Global regulations are developing at pace with territories taking a variety of approaches ranging from voluntary to mandatory and covering specific sectors or taking an economy-wide approach.

[80] https://www.enisa.europa.eu/publications/artificial-intelligence-and-cyber security-research

Conclusions

The potential benefits are there to be leveraged for the good of humankind. AI can deliver benefits across science and medicine, agriculture and the environment, creative arts and professional services. Autonomous systems will be able to operate in environments not safe for humans and automation will increase productivity, and potentially raise living standards in the process. However, as with any technological revolution, there is a risk that sections of society will be left behind, increasing and entrenching inequality.

To bring it back to AI and its applications in cyber security, there is no panacea. Organisations should:

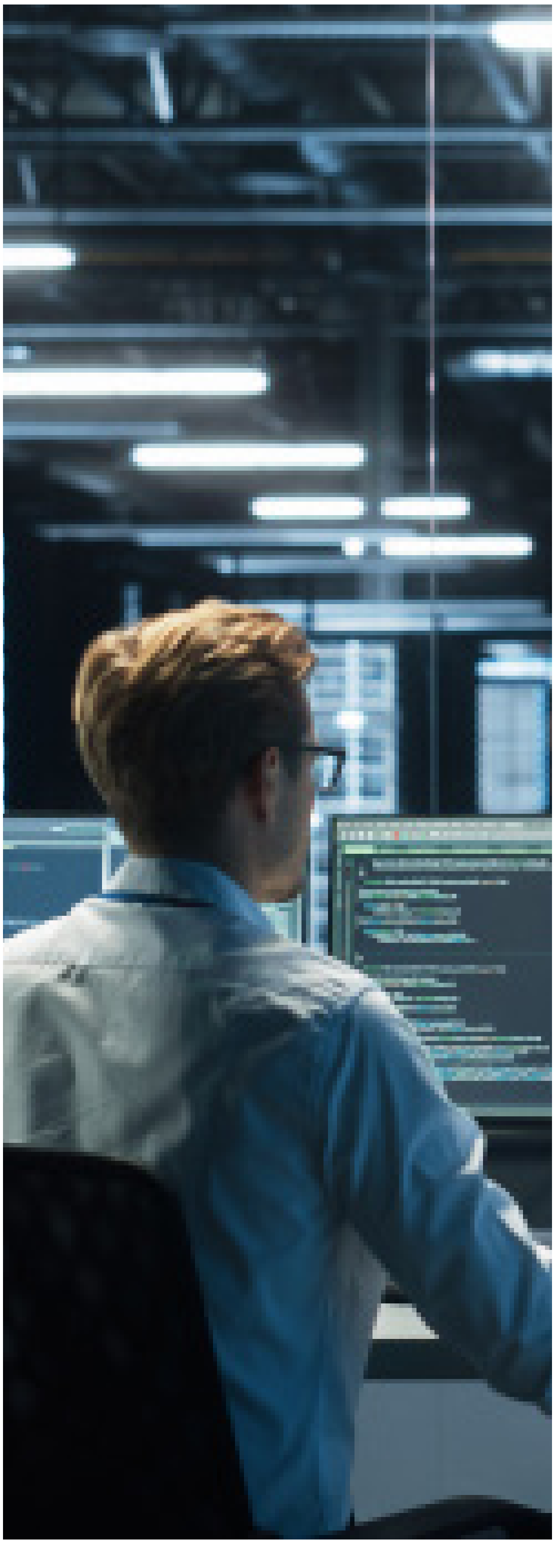
- Understand their threats, manage risks and seize the opportunities presented by AI to improve cyber posture.

- Adopt a hacker mindset by conducting penetration testing and red teaming against AI systems, and with the presumption that attackers are leveraging AI to improve the efficacy of their attacks.

- Ensure that privacy, information security and ethics, as well as evolving regulatory requirements, are considered when developing or using AI within business applications, and that AI systems and applications are developed in line with relevant regulatory requirements

- Adapt safety processes to ensure that risks in safety critical systems utilising AI are understood and managed.

Finally, cyber security leaders need to communicate with employees and educate them about appropriate use, and with their boards and executive teams to explain the risks and opportunities posed by AI and how they are being managed.



About the Authors

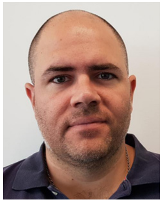
Verona Hulse, UK Head of Government Affairs

Verona is an experienced government affairs and policy professional currently leading UK public affairs for FTSE250 global cyber security firm NCC Group. In this role, she oversees NCC Group's engagement with UK government and regulatory decision-makers and the wider policymaking community, against a backdrop of the increasing regulation of cyber resilience. Prior to joining NCC Group, she has overseen in-house and consultancy public affairs programmes for a range of organisations – from FTSE100 and public sector institutions to start-ups and disruptors, across many sectors of the economy including aviation, logistics, and utilities.

Jon Renshaw CEng MIET, Deputy Director of Commercial Research

Jon has consulted on the design and implementation of cyber secure communication systems across the public sector, defence, transport and manufacturing industries. He began his career in cyber security at Airbus Defence and Space after completing a bachelor's degree in electrical and electronic engineering at Cardiff University. He subsequently took on a senior engineering role at a network security consultancy startup, ITSUS Consulting, where he managed the company's innovation processes before moving on to setup his own consultancy business. Most recently he has worked as a secure network architect for Thales Cyber Defence Solutions before joining NCC Group in 2022 as Deputy Director of Commercial Research where he is responsible for delivering fundamental and applied cyber security research in collaboration with, and on behalf of, NCC Group's customers.

Jon is a Chartered Engineer (CEng) with the UK Institute of Engineering and Technology (IET) and has achieved qualifications in enterprise architecture, information security risk management, cloud technologies, organisational leadership & management.



Jose Selvi, Executive Principal Security Consultant

Jose has worked in the cyber security industry since 2004. During this time, he has had the opportunity to work in different fields such as Intrusion Detection, Incident Response, APT hunting & intelligence, and Computer Forensics. However, offensive security, penetration testing, and security research are the fields where he has focused most of his career. Although Jose is accustomed to testing all kinds of software and devices, in the last few years, he has been involved in the AI Security field. He has published several blog posts on this topic, in which he also holds a PhD.

In addition, Jose is always eager to contribute to the infosec community by releasing tools and presenting at security conferences. DEFCON, BlackHat EU, Ekoparty, OWASP Appsec, and SOURCE Conferences are examples of well-known conferences where he demonstrated practical attacks against NTP (Delorean) and a browser side-channel (FIESTA) that were used to break TLS security.

Our global experts in AI and cyber security

Matt Lewis, Global Research Director

Matt is a Technical Research Director with over 22 years' experience in cyber security. His specialisms include general security consultancy, scenario-based penetration testing, vulnerability research on new and emerging technologies, and development of security testing tools. His penetration testing, research, and consultancy experience spans multiple technologies across all sectors, including many FTSE 100 and Forbes 2000 companies. Matt is also an experienced public speaker and contributes his expertise on a wide range of cyber security topics at global conferences and various print and media outlets.

Chris Anley, Chief Scientist

Chris has been carrying out security audits since 1996, performing thousands of penetration tests, code reviews and design reviews on a wide variety of platforms, languages and architectures for many of the world's largest companies. As Chief Scientist, he promotes, advises and assists with NCC Group research programs, as well as carrying out independent research into new and emerging security threats.

He has published a wealth of security research in a variety of areas including foundational papers in application security, several books on database and application security and a range of research into Artificial Intelligence and Machine Learning.

Gerben van der Lei

Gerben has been with NCC Group for 11 years. Holding a degree in applied physics, he currently serves as a Strategy Program Manager. Areas of interest and expertise are AI, cryptography, and both software and hardware development. Previously Gerben was responsible for managing NCC Group's portfolio of government-grade cryptography products providing unique experience in the full security life-cycle. He has extensive expertise in data analysis and data-driven optimizations.

Our global experts in AI and cyber security

Eric Schorn, Technical Director

Eric is a Technical Director on the Cryptography Services team at NCC Group. He writes, reviews and tests both AI and cryptography-related software applications. He has led large projects ranging from cryptographic primitives involving BLS Signatures in assembly language, through to complete business systems such as VPNs and encrypted backup solutions in C++, Golang, Rust, Python and even Erlang. Eric has 25+ years of industry experience and 14 US patents issued.

Liz James, Managing Security Consultant

Dr. Liz James holds a Doctor of Engineering degree and currently serves as a Managing Security Consultant at NCC Group. With a research background focused on On-Board and Off-Board Data Platforms, Liz specializes in designing and implementing proof of concept solutions for complex embedded systems, particularly in the automotive sector. Her doctoral studies at WMG, University of Warwick, from 2015 to 2019 have sharpened her expertise in these areas.

Since joining NCC Group in 2020 as a Security Consultant and subsequently being promoted to Senior Security Consultant in 2022, Liz has been instrumental in designing robust security architectures and conducting comprehensive vulnerability assessments. Her work includes conducting penetration testing on various vehicles, entire systems, individual components, and IoT devices. With a strong focus on the design and implementation phases of customer projects, Liz strives to ensure the safety, security, and resiliency of intricate systems.

Liz's dedication to the field extends beyond her role at NCC Group. She actively contributes to the technology sector as the Vice-Chair of the Intelligent Mobility and Transport Steering Board at techUK, further demonstrating her commitment to advancing intelligent mobility solutions. With her extensive knowledge in secure system design and testing, coupled with her leadership position, Liz continues to make significant contributions to the field of technology and security.

Thomas Atkinson, Managing Security Consultant

Thomas has been with NCC Group for 7 years. He comes from a scientific research background and is a technical managing security consultant. His specialisation is in all thing AI/ML and the majority of his work is client paid for research around hacking or using AI/ML in a security related context.

Acknowledgements and NCC Group Overview

Acknowledgements

Thank you to all the NCC Group researchers and blog authors who continue to develop interesting and insightful articles on the application of AI to cyber security and vice versa. And thank you to all the reviewers who generously shared their time to ensure this paper is coherent and impactful.

About NCC Group

NCC Group is a global cyber and software resilience business operating across multiple sectors, geographies and technologies.

What We Do

We assess, develop and manage cyber threats across our increasingly connected society. We advise global technology, manufacturers, financial institutions, critical national infrastructure providers, retailers and governments on the best way to keep businesses, software and personal data safe.

About NCC Group Research

NCC Group supports its consultants to carry out exploratory research into topics across the cyber security landscape. Many of these research endeavours result in blog posts published on our website (https://research.nccgroup.com) and as presentations at conferences around the world.

We also support our customers to answer their research questions through our commercial research service which covers horizon scanning, proof of concept development, vulnerability and control efficacy research and collaborative and consortia research.

Disclaimer – All content in this whitepaper was created by human employees of NCC Group, except where explicitly stated otherwise.

Further Reading

Technology Overview

Machine Learning algorithms require significant specialised computing resources for optimal efficiency, especially during the learning phases of building a model. Since Machine Learning algorithms often require intense computations, GPUs provide much higher performance for certain types of operations than Central Processing Units (CPU). This is due to their ability to run massively parallel computations across many cores versus a CPU which processes tasks in a much less parallel, but flexible, fashion across just a handful of cores in comparison.

As model sizes, and the amount of data they are trained on, have increased in size it becomes inefficient to train them using a single server. Clusters of servers, each containing multiple GPUs, are often networked together with learning operations distributed across them.

As well as the computations required to train the models, the data they are trained on requires specialised "big data" storage systems to provide seamless access to large unstructured and semi-structured datasets.

The variety of use cases these models are deployed in is growing to include autonomous systems. These systems support a variety of sensors to detect the state of their environment and based on this state, in combination with a goal, can determine the best action and to effect this through actuators. A well-known example of this is a self-driving vehicle, it is able to sense its location and surroundings (including other road users and signage) through a combination of cameras, lasers and radar. Based on its destination it can determine a route and accelerate, brake and steer to reach its intended location. But any AI system where the outputs have a physical effect in the real world has safety concerns, these are described in further detail in Safety.

Although AI has novel features which bring with it unique cyber security challenges and opportunities, it is still a combination of software and data, processed and stored on hardware, provided and possibly managed by a myriad of suppliers. All of the existing concerns about securing your data and systems still apply in an AI context.

[81] https://www.deepmind.com/research/highlighted-research/alphago

Further Reading

Glossary of Terms Relevant to AI

Artificial Intelligence (AI) and Machine Learning (ML):

These two terms are frequently synonymous, although ML is a more precise term for most technologies we use currently. Machine Learning, as its name indicates, is the technique used for computers to learn functions and solve tasks based purely on data, with no additional human instructions. The nature of the functions and the techniques to adjust it to the problem may differ depending on the model in use, but all of them essentially perform this function.

Supervised vs Unsupervised Learning:

Data collected for training an ML model may include the expected output for a given input, or just inputs. That is the fundamental difference between supervised and unsupervised learning. In supervised learning, the difference between expected outputs and predicted outputs can be used to adjust the model function and improve further predictions. Unsupervised learning uses a different set of techniques to adjust those functions with no additional information, such as in "clustering".

Reinforcement Learning:

This is an alternative approach that is not actually relying on data itself, but on the interactions between the model and an external environment. For this purpose, a reward function is created, so a more beneficial interaction will result in a higher reward, and the learning process will aim to maximise the overall reward. Although this is the classic approach for gaming[81], it has also been used in language models for fine tuning.

Tasks and Problems:

Models can solve several different problems, depending on how they are designed and trained. The most well-known problems are:

Regression: They find functions that predicts output values corresponding to certain inputs. The classic example
is predicting residential property pricing based on characteristics such as the location, size, crime rates etc.

Classification: This problem requires functions to split the training dataset so all the inputs with the same expected classification are in the same region. Outputs predict the class the input belongs to. For example, given a vector
of features of a binary file, classify if it is malware or benign, or even detect which family of malware is it.

Clustering: Similar to classification but, since no expected classes are included in the training dataset (unsupervised learning), the training groups together those inputs that are more similar to each other. These techniques are also closely related to outlier/anomaly detection.

Generation: This problem requires generating new content (typically text or multimedia content) based on a given input. Some of the most famous AI models currently are generative models, such as most Large Language Models (LLMs) which are based on the transformer architecture[82]. Training these models usually requires hiding part of the input data (certain words or parts of the multimedia content) and finding a function that predicts the hidden element.

[82] "Attention Is All You Need" https://arxiv.org/abs/1706.03762

Further Reading

Glossary of Terms Relevant to AI

Hardware Requirements:

Although there are very simple ML models that can be trained and executed on low resources equipment, models solving complex problems and in particular deep learning models require much more powerful equipment, costing a minimum of tens of thousands of dollars. The fundamental components are Graphical Processing Units (GPU) which are specifically designed to perform high performance matrix multiplications. These calculations are a fundamental requirement for deep learning, especially training. Recently, the new generation LLMs have been found to be much larger than any other models seen before. For that reason, GPUs with large memory specifications are required.

Software Requirements:

Several open-source platforms such as scikit-learn[83], TensorFlow[84] and PyTorch[85] are frequently used to implement ML models. In addition, communities such as Hugging Face[86] provides an extensive platform to implement and share models. This enables users to download pre-trained models across many domains and use them in an easier and faster way than implementing and training them from scratch.

AI as a Service (AIaaS):

Outsourcing the management of the hardware, operating systems and software used to develop and/or run AI services lowers the barrier to entry to those wishing to experiment with or leverage the benefits of AI. Cloud service providers supply a variety of offerings ranging from hardware focused Infrastructure as a Service (IaaS) tailored to the requirements of AI workloads up to Software as a Service (SaaS) chat-bots and API interfaces to ready-made or bring your own models.

[83] https://scikit-learn.org/stable/
[84] https://www.tensorflow.org/
[85] https://pytorch.org/
[86] https://huggingface.co/

About us

People powered, tech-enabled, Cyber Security

NCC Group is a global cyber business, operating across multiple sectors and geographies.

We're a research-led organisation, recognised for our technical depth and breadth; combining insight, innovation, and intelligence to create maximum value for our customers.

As society's dependence on connectivity and the associated technologies increases, we help organisations to assess, develop and manage their cyber resilience posture to confidently take advantage of the opportunities that sustain their business growth.

With over 2,400 colleagues, we have a significant market presence in the UK, Europe and North America, and a growing footprint in Asia Pacific.

Contact Us:
+44 (0) 161 209 5200
www.nccgroup.com

XYZ Building
2 Hardman Boulevard
Spinningfields
Manchester
M3 3AQ