LLOYD'S

Generative AI: Transforming the cyber landscape

March 2024

Lloyd's Futureset

Executive summary

Major leaps in the effectiveness of Generative AI (GenAI) and Large Language Models (LLMs) have dominated the discussion around artificial intelligence over the past 18 months. Given its growing availability and sophistication, the technology will inevitably reshape the cyber risk landscape.

This report explores how GenAI could be used by threat actors and cyber security professionals and highlights its potential impacts on cyber risk. Below we summarise the LLM landscape, the transformation of cyber risk, the considerations for business and insurance and the ways in which Lloyd's will take action to develop solutions that build greater cyber resilience.

1. The LLM landscape

Generative AI and LLMs (Large Language Models) are a very new set of technologies, with pivotal enabling advancements happening only about 6 years ago.

Major leaps in model effectiveness across a variety of tasks relevant to cyber security have occurred in the last 18 months and are likely to continue into the near future.

Applications of LLMs to cybercrime have been minimal to date due to effectiveness of AI model governance, cost and hardware barriers, and content safeguards.

The release of unrestricted frontier models plus recent algorithmic efficiency discoveries represent a pivotal breakdown in AI governance. There are now many publicly available models which can create explicitly harmful content, and they can now be run on commodity hardware cheaply.

2. Transformation of cyber risk

Vulnerability discovery: Automated vulnerability discovery, especially in domains which elude human experts, is likely to significantly increase the pool of options for threat actors. Threat actor tooling is likely to outpace defensive tools created by the security industry due to asymmetric incentives.

Campaign planning and execution: Cyber-campaign targeting and scoping is likely to become cheaper, more fine-tuned, and broader due to automation of target discovery. This would mean threat actors could generate bespoke attack materials for many potential targets.

Risk-reward analysis: Threat actors' ability to evade attribution and achieve their desired outcomes (exfiltration of funds, etc) is likely to be enhanced. This could shift risk-reward calculations in their favour and embolden them.

Single points of failure: The rise of a new class of service provider linked to the provision of LLMs, could generate a new type of single points of failure. Losses arising from interruption or compromise of these single points of failure are likely to be different from what we expect today.

3. Considerations for business and insurance

A new threat landscape: AI is likely to augment threat actor capability, enhancing the effectiveness of skilled actors, improving the attractiveness of the unit cost economics, and lowering the barrier to entry.

Cyber catastrophes: There may be a modest increase in the risk of manageable cyber catastrophes. In contrast, smaller scale events are likely to increase at a greater pace as AI-enhancements allow threat actors to more effectively design targeted and lower profile campaigns.

State-backed, hostile cyber activity: AI has the potential to improve the effectiveness of state-sponsored hostile activity, both in terms of espionage and sabotage. However, it is unclear to what extent the proliferation of advanced capabilities will increase the risk of a major catastrophe happening, due to the human factor.

4. Taking action

At Lloyd's, we will continue to work with our stakeholders to support the development of this important technology, and evaluate and address the potential threats.

Educating our community: Build awareness and education through Lloyd's Futureset as the growth of AI technology transforms the risk landscape.

Partnering with industry: Encourage greater collaboration with governments, regulators and technology companies to better manage AI risks.

Engaging policymakers: Work with insurers, governments and others to inform the development of intelligent policy guiderails that can support this important technology.

Supporting sustainable innovation: Enable new product development through the Lloyd's Lab that will serve as a springboard to develop new solutions responding to the changing cyber threat landscape.

Introduction

Discussion regarding the impact of Artificial Intelligence (AI) on society has been dominated by the emergence of Large Language Models (LLMs) into public awareness over the last 18 months.

While there have been many advances in adjacent research areas, none have captivated the public's imagination or anxieties to the same degree as generative models in textual and graphical domains. Due to the rapidity of advances in AI research and the nature of the highly dynamic cyber environment, analysis of the consequences these tools may have for cyber perils has been limited.

The goal of this report is to provide context and grounding around how generative AI models ('Gen AI') have emerged and the safety mechanisms that have so far prevented their widespread misuse by threat actors. It goes on to discuss the implications of recent events that have resulted in a decoupling of this technology from these safety mechanisms and analyse how the cyber risk landscape could be transformed by its proliferation.

Lloyd's has been exploring the complex and varied risks associated with AI since developing the world's first autonomous vehicle insurance in 2016. We're all early on in exploring the world of AI, its risks and opportunities – but our market is designed to bring expertise together to underwrite the unknown.

AI is already being used in our market, such as Tautona's claims processing chatbot and Reor20's AI-based flood loss modelling, both developed in the Lloyd's Lab. We see significant opportunities for AI to make life easier for our customers and those using our market, but also substantial risks in the underwriting of AI – where the field continues to change every day.

Lloyd's is committed to working with insurers, startups, governments and others to develop innovative products and intelligent policy guiderails that can support the development of this important technology and create a more resilient society.

The concluding sections of the report explore the implications of Gen AI on cyber risk for businesses and the insurance industry, along with the steps that Lloyd's will take to support our customers and market in responding to the evolving and highly unpredictable threat; working with governments, regulators, risk experts and insurers to pioneer new ideas and support the growth of resilient cyber solutions.

As the world's leading marketplace for commercial, corporate and specialty risk insurance and reinsurance, Lloyd's is committed to building resilience against cyber risk.

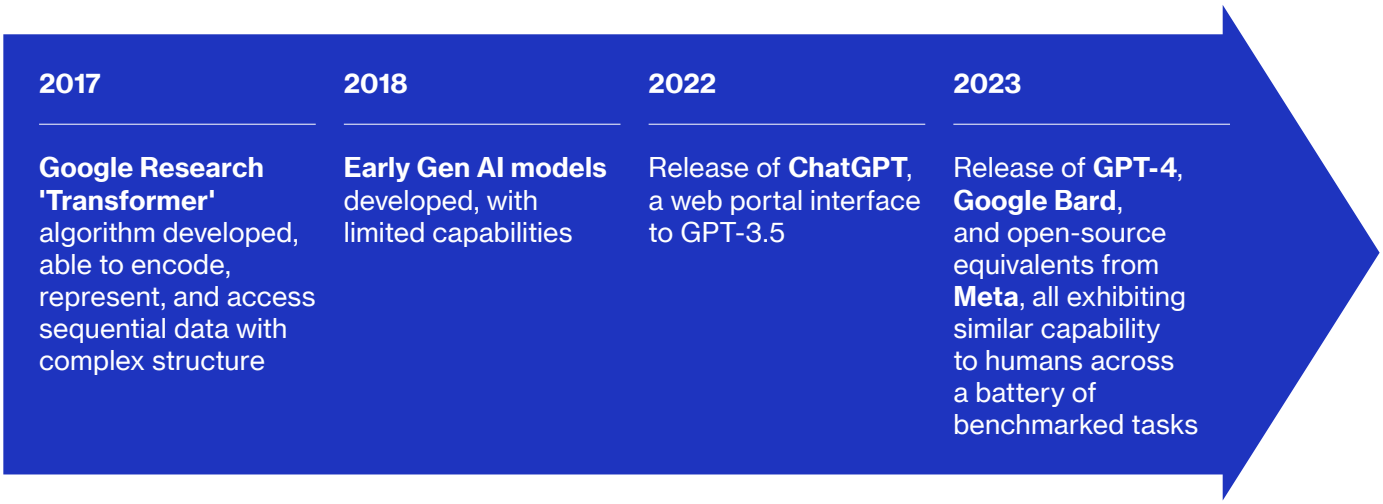
AI and other emerging technologies present both opportunities and risks for businesses and the insurance industry. We like to keep an eye on the positive – while doing what we can to mitigate the risks – which is why Lloyd's is committed to exploring and responding to the implications of these rapidly changing technologies.

Dr Kirsten Mitchell-Wallace
Director of Portfolio Risk Management

The story so far

LLMs: a brief history

Approximately 6 years ago, a seminal paper[0] was published by Google Research introducing a novel algorithm for encoding, representing, and accessing sequential data with complex structure. This machine, dubbed 'Transformer', would underpin almost all language, vision, and audio based generative machine learning approaches by 2023.



The primary contribution of this work was to enable processing of vastly larger amounts of data than was feasible previously on a fixed computational budget. This was achieved by transforming a sequential processing problem, where operations were performed in order one at a time, into a parallel one, enabling multiple steps to be performed at the same time. This breakthrough reduced the computation cost of the algorithm and unlocked the ability to model long-range structure in data. Subsequent research has continued in the same way: driving down computational cost, improving the granularity of the information representable by the models, and increasing the size of the input data corpus.

Early models (circa 2018) were limited in their capability, but progressively scaling up the computing power and dataset size resulted in rapid advances, culminating in the release of ChatGPT, a web portal interface to GPT-3.5, which was released less than 18 months ago (Nov '22), to considerable public interest and concern. Since November, notable events include the release of GPT-4 (March '23) - exhibiting similar capability to humans across a battery of benchmarked tasks, Google's Bard (March '23) and completely open-source equivalents by Meta (March, July '23).

The velocity of major advances in this field has created difficulties for enterprises seeking to enact reasonable AI Governance policies, as well as regulators and policy makers seeking to ensure the evolution of the technology proceeds in the public's interest. A keen focus on 'AI Safety' by research labs has been successful in preventing the widespread proliferation and misuse of this technology until very recently. Differing approaches to regulation have emerged, with the EU AI Act taking a risk based, top-down legislative approach, the UK choosing a white paper driven, principles approach[1,2], and the US taking a nonbinding, soft-law approach to governance which takes care to avoid disrupting investment and growth in the US technology sector[2]. The commercial sector has responded with lobbying attempting to weaken regulation, even while simultaneously calling for stronger AI safety guardrails[3].

Safety: AI model governance, financial barriers, and guard-rails

The rise of powerful generative models brings with it tremendous opportunity for innovation, but also introduces significant risks of harm and misuse. It is fair to ask why despite the claims of advanced capabilities of these tools, few material impacts on the cyber threat landscape seem to have occurred. The answer has so far been that the industry focus on safety has prevented widespread misuse, as well as economic considerations.

"AI Safety" is a term without a consensus definition, referring to several related and interlinked areas, which can be classified in three broad categories:

A Autonomy and advanced capabilities – calls for oversight and control of "systems which could pose significant risks to public safety and global security"[4,5]

B Content generated by the models – potentially leading to issues with privacy, copyright, bias, disinformation, public over-reliance, and maybe more

C Malicious use of the models – leading to harm or damage for people, property, tangible and intangible assets

The first sense of AI Safety **(A)** has received increasing attention in 2023, with governments allocating resources to understanding the risks involved. The UK government has created a 'Frontier AI Taskforce' consisting of globally recognised industry experts, tasked with advising and shaping policy pertaining to the creation of powerful models. However, despite the growth in interest and investment, the nature and extent of the risk posed by these systems is still unclear. Due to the lack of quantitative or even qualitative information on this topic, it will not be considered further in this report, but is an area which it will be important to monitor as the situation develops.

Research labs, commercial enterprises, and policymakers have focused on understanding safety in the sense of **(B)**, which related mainly to issues around bias, privacy, fairness, and transparency. All serious issues, that are rightly being the subject of active research, and that also have potential consequences for policies beyond Cyber or Tech E&O, but that are however beyond the scope of this report.

The remaining pressing concern is the question linked to **(C)**: "How can the risk of harm or damage arising from human actors intentionally using these models maliciously be mitigated?"

Broadly, there are three mechanisms which have underpinned the safety apparatus curtailing malicious use of generative AI technology. The effectiveness and roles in mitigating the risk of cyber threat will be briefly discussed for each, after which a recent pivotal event causing a breakdown of these mechanisms will be examined, along with impending consequences for cyber risk:

1. AI model governance

Managing public access to critical trained model components and artifacts, ensuring and promoting model safety as part of broader AI safety goals.

Key elements of model governance for enterprises and research groups producing LLMs include:

- Output artifacts of the model training processes (known as model 'weights') are kept secret and not released to the public. Possessing only model code and having access to the computing hardware is insufficient to run the models. The weights of these models are kept as closed as possible to create commercial and regulatory moat, and to prevent misuse
- Model training and inference (serving requests) takes place on private computing infrastructure, with internal details opaque to end users
- Setting and following rules for monitoring AI models in areas like quality, performance, reliability, security, privacy, ethics, and accountability
- Application of governance principles throughout the entire lifecycle of AI models: training, analysis, release (if applicable), deprecation

The key outcome of model governance is preventing the public from having oversight-free access to emerging disruptive technologies until adequate safety controls can be enacted: technological, regulatory, legal, or otherwise.

2. Financial and computational barriers

Costs or hardware requirements for training and running large models.

The process of training, fine-tuning, or performing inference with large generative models is computationally intensive, requiring specialised computing hardware components. For context, training a recent model released by Meta (Llama2) required "3.3M hours of computation"[6], with total training costs estimated around $10M for electricity and usage of the hardware – this figure does not consider the acquisition and construction of the data centre itself, or the cost of staff.

Inference tasks on these models have less exorbitant requirements but still have until recently required access to a datacentre, with prohibitive costs for most threat actors. However, recent developments have driven these costs down, as will be discussed in the next section.

The consequences of this have been the inability of the public, including small research labs or universities, to train or run their own large models, restricting them to much less capable versions. All access to 'frontier-grade' generative models has been through the large labs (OpenAI, Meta, Anthropic, Google), and is subject to their strict governance, oversight, and safeguards.

3. Safety guard-rails

Safety fine-tuning, access control, and usage monitoring of generative AI models and tooling.

An important consequence of controlling the training process of LLMs and restricting public access to the models through custom interfaces is the ability to apply strict controls to their usage in accordance with the governance safety principles of the hosting organisation.

All large commercial models to-date except for those released by Meta have ensured access was closely safe-guarded and monitored through specialised interfaces like ChatGPT or Bing Chat. Commercial state-of-the-art LLM training involves a safety pipeline with several key elements:

- Strict curation of the input training data to ensure minimal toxic, illegal, or harmful content can be seen by the model
- Specialised 'fine-tuning' techniques involving human curators who provide feedback on potential model responses (incl. refusals to produce harmful content)
- Adversarial testing with domain experts assessing capability of model, especially with respect to emergent behaviours
- Layers of evaluations and mitigations ensuring the models adhere to safety principles
- Strict curation of user-interface access: requests and responses are screened and logged via web interfaces, dangerous requests result in revocation of access and potentially local authorities

If users do not have full access to the models and all internal components, it is impossible to circumvent these restrictions in any meaningful way; while some 'jailbreak prompts' may allow a soft bypass, it is ineffective for very harmful requests. Likewise, users cannot bypass screening mechanisms if forced to interact with the models through online portals.

The cloud-hosted solutions ChatGPT, Bing, and Google's PaLM are examples of LLMs with extensive safety controls on them preventing misuse, and much of the public's concern about the potential to use these models for harmful or illegal purposes has been mitigated by virtue of these controls.

Opening Pandora's Box: Proliferation of zero-safety LLMs

In February 2023, Meta's AI Research group announced their internal model Llama, and shortly after, the internal model 'weights', alongside the code, were leaked to the public. Several months later in July, Meta released a vastly more powerful model, Llama2, to the public openly under a commercial license in its entirety, including the model weights.

This was the first time a state-of-the-art model was available in its entirety to the public and initiated a storm of development activity. On a popular public model repository, there are hundreds of freely available variants and derivatives of the Llama1/2 base models, including ones which have had all internal safety mechanisms stripped out, or have been fine-tuned on explicitly malicious or harmful content.

Due to the effectiveness of the safety approaches discussed in the previous section, the capability of 'unaligned' models to produce harmful content is underappreciated. As stated in the GPT-4 Technical Whitepaper, some example tasks the unaligned base model could perform[7,8] as long ago as 2022:

- Writing detailed, accurate explanations on how to conduct money laundering activities with reference to dark-web locations for key elements of the process.
- Analysing code and devising ways to exploit vulnerabilities in it.
- Writing compelling and harmful misinformation targeting individuals or groups of people
- Writing phishing emails or producing direct communications convincing individuals to take harmful actions.
- Impersonating a human well enough to convince a remote worker to perform a task allowing the bypass of a security control[1]

In the months since its release, new techniques and approaches for enhancing LLM model capabilities in narrow areas have been discovered, and it is a certainty that providing high-quality training data input for malicious tasks will augment model capabilities further.

[1] As part of safety research, GPT-4 was tasked with gaining access to a page protected by a CAPTCHA. Having no way to 'see' the CAPTCHA but able to interact with the browser, the model devised a plan wherein it contacted a worker in a freelance market (TaskRabbit), convinced the worker it had a vision disability preventing the completion of the captcha and needed assistance completing it, shared the image, and successfully gained access to the target page. This demonstrates complex problem-solving skills, as well as an understanding of human psychology.

Driving down costs

Additionally, since February, several advanced techniques have been developed which have dramatically driven down the computational requirements for training, fine-tuning, and running inference for these models. Hundreds of LLMs now exist in the wild for a variety of tasks, many of which can be run locally on commodity hardware.

As of September 2023, it is possible to run a LLM with capability equivalent to GPT3.5 on a consumer grade hardware such as a MacBook M2, completely locally (without internet connection). This means that all safeguards detailed above can be completely circumvented: models can be adjusted to answer all requests regardless of harm, and this can be done in completely sealed, cheap local computing environments, without any oversight.

We are entering a period where no meaningful safeguards or harm-reduction curation through centralised ownership and management of LLMs will be applicable to threat actors – an era of proliferation. Powerful, specialised-purpose models will be easily created, distributed, and run on commodity hardware for the purposes of cyber-crime. It will take some time until the extent of the capabilities for illegal purposes are understood and industrialised, but this work is certainly already underway by threat actors. While this has significant implications for the threat landscape, further harm can be mitigated by preventing the release of future advanced models or proliferation-enhancing technologies through regulation.

Transformation of cyber risk

Framework: drivers of cyber threat

The following illustrative framework explores how Gen AI tools could be used by threat actors or cyber security professionals and highlights some potential impacts on cyber risk. While there is no definitive list of components which drive or enable the formation of cyber crime campaigns, it is possible to consider the following factors which influence the frequency and severity of cyber threats in predictable ways and assess the potential impact of emerging LLM technology on each of them.

Drivers of cyber threat		Evidence	Potential impact
01. Vulnerability discovery	The deeper and more diverse the pool of vulnerabilities available to a threat actor, the more options they have for approaches, the lower the cost of creating exploits, and the lower the opportunity cost to execute a campaign.	High	Very high
02. Campaign planning and execution	Campaigns require the ability to specify and describe a target group, understand their technology usage, operational behaviours, security posture, and willingness to pay. Materials required for the execution of the campaign need to be created for the identified targets, taking time and resources for the threat actor.	High	High
03. Risk-reward analysis	All criminal activities which are not terrorism or war-like involve an assessment of risk and reward. Modifying the effectiveness of mechanisms for obtaining illicit gains or evading law enforcement can shift the equation in predictable ways.	Low	Low/Medium
04. Single points of failure	A fourth component underpinning the above three is the degree to which systems, services, technologies, and people are bound together (systemic coupling), giving rise to single points of failure. The greater the systemic coupling, the greater the unit-cost effectiveness of vulnerabilities, the larger the exposure footprint for any set of exploits, and the more extreme the risk-reward factors become. Examples of systemic coupling are cloud providers, DNS providers, and typically all services used by many businesses and provided by a small number of firms.	Low	Medium/High

Very high High Medium/High Low/Medium Low

1. Vulnerability discovery

The existence of software or hardware vulnerabilities to exploit is a requirement for almost all forms of threat actor action, except for purely social engineering-based approaches. It is time consuming to analyse code bases for significant vulnerabilities which can lead to exploits and requires significant expertise.

Gen AI enabled evolution	Automation of vulnerability discovery
Evidence	**High** Significant progress has already been reported in this area, and open source tools are available today. Specialised, restriction-free LLM performance benchmarks suggest they will excel at this task, particularly for 'hard' domains.
Potential impact	**Very high** Automated discovery of vulnerabilities may increase frequency of all types of cyber loss significantly, due to the expanded pool of options, and associated downward pressure on the cost of purchasing exploits. Enhanced tooling can increase cost efficiencies for both threat actors and security professionals; potentially asymmetrically. The potential of vulnerabilities discovered in esoteric industrial control system software raises the possibility of cyber-physical risks as well.

Usage of LLMs fine-tuned for code analysis to identify exploitable programming errors has the potential to drive the 'cost-per-vulnerability' down by orders of magnitude relative to human investigation by performing at-scale scans of open-source repositories.

- LLMs enable automated vulnerability discovery, even in domains which are very challenging for humans, such as:
  - Embedded micro-code and firmware
  - Decompiled proprietary binaries in closed source enterprise software
  - Hardware device drivers

A larger pool of vulnerabilities to choose from grants a greater flexibility in design of exploits, choice in targets, and campaign methodologies. Successful attacks require a series of vulnerabilities to be exploited, allowing attackers to progressively gain initial access to their target, create a foothold and traverse the systems, and finally create an impact. Regardless of the specifics of these factors, having more surface in a target system makes every step easier, cheaper, and more economical.

- AI-enhanced dynamic analysis tools for discovering vulnerabilities in 'live' software environments or networks could be a force multiplier for very skilled actors
- LLM-powered malware can run on physically unobtrusive, portable hardware like a Raspberry Pi device. While capabilities of malware on low-power devices will be more limited, these tools could significantly increase the risk associated with physical access vectors

While security professionals and vendors will likely look to utilise similar LLM-enhanced tooling defensively in areas like threat intelligence, incident response, and monitoring and detection, there remain fundamental asymmetries that may provide threat actors an advantage.

Threat actors and their organisations will likely have greater incentives and flexibility to construct highly customised tools for narrowly focused augmentation tasks. The potential financial rewards for novel vulnerability discovery or exploitation provide strong motivation to explore even obscure and highly specialised targets, and the risks involved with their activities result in an extremely high cost of failure.

Small groups and individuals can also allocate time and resources without oversight, whereas security teams in enterprises face organisational constraints and may not be strongly incentivised to create defences against low-likelihood attacks in addition to being less agile. Additionally, legacy software in IT and operational technology (OT) contexts may be difficult or impossible to patch or defend, providing attackers with a large surface area to exploit.

In general, the decentralised nature of the adversary ecosystem would allow for more experimentation and specialisation compared to the more homogeneous security teams within centralised enterprises, and with stronger incentives.

2. Campaign planning and execution

Gen AI enabled evolution	Automation and enhancement of many elements of campaign planning and synthesis of materials required for execution
Evidence	**High** Models already exist that are capable of passing as humans in restricted contexts and can perform data collection, analysis, and planning tasks. Gen AI content synthesis is already creating significant issues (such as 'deepfakes'), with several high-profile enterprise breaches occurring due to executive video impersonation. LLMs have demonstrated the ability to construct convincing phishing materials, as well as extortion or blackmail attempts.
Potential impact	**High** Cyber campaign effectiveness is likely to increase significantly due to reductions in planning and preparation cost-per-target, suggesting we will see broader, more frequent campaigns with more severe losses. Attacks with the aims of Data Breach/ Destruction, Extortion, or Reputational Damage may be strengthened by the increased difficulty of determining the authenticity of communications from key leaders within organisations.

A critical factor determining the breadth and severity of a cyber campaign is the balance between the following:

- Ability of the threat actor to survey for potential targets, identify their characteristics, collect data relevant to the attack, and produce materials required for attack execution
- Resources available (human resource, financial, political) to conduct the investigation and preparation for the campaign

Phishing campaigns, whether broad or targeted, require significant amounts of human attention to design and execute, and this is a limiting factor in their cost-effectiveness. Spear-phishing requires an in-depth profile of the target, collected through investigation, and potentially even a human actor-in-the-loop to follow through with subsequent steps of the attack. Data breach campaigns require careful selection of targets based on the sensitivity of data held, public sentiment about a breach, and judgements about willingness to pay.

Generative AI offers the possibility to automate and enhance many elements of the campaign targeting chain:

- Data collection on target individuals or companies, and profile analysis, shortlisting, and prioritising of targets via mass data scraping and advanced analysis by LLM
- Complete automated synthesis of phishing, impersonation, defamatory or blackmail attack materials: textual, audio, or visual content. This will enable threat actors to impersonate leaders and financial officers and use fraudulent communications to enable access to an organisation's networks, communications, and sensitive information. The NSA has released an in-depth advisory on the risks[9]
- High quality translation of campaign materials into other languages to reach a broader audience
- LLM real-time interaction with campaign targets replacing need for human threat actors

3. Risk-reward analysis

Gen AI enabled evolution	A deeper ability to conceal attacks, and reach a larger number of targets undetected
Evidence	**Low** It is unclear to what extent LLMs will be able to assist with evasion of fingerprinting and attribution, or with safe exfiltration or laundering of funds. Cryptocurrencies have enabled financial motivations in the form of ransomware, but to date there have been limited intersections between crypto and AI.
Potential impact	**Low/Medium** Lack of attribution could empower threat actors by making it difficult to model their behaviour, allocate law enforcement resources effectively, and track changes over time. Overall, it could decrease confidence in our risk assessment of the cyber threat landscape and may increase the frequency of threat actor campaigns.

All rational threat actor activities involve some kind of risk-reward analysis. A rational threat actor will require a means by which they can extract some form of utility for their activities, whether that be money, information, reputation, political goals, or otherwise. Critically, it is in the interest of all threat actors to execute their campaign without attracting too much attention from powerful adversaries (such as state intelligence agencies), constraining the scope of the attack to only the most necessary targets. Methods to obscure the 'fingerprints' of the actor, to obfuscate the components of their attack, or to exfiltrate funds through complicated chains of intermediaries are common.

LLMs could potentially enhance these obfuscation activities through automated scrubbing of identifying characteristics of malware, cyber campaigns, or through intentional fingerprint misdirection.

4. Single points of failure

Gen AI enabled evolution	A small number of providers of reliable Gen AI models, and the sharing of training data across large swathes of the business population.
Evidence	**Medium** There is evidence of concentration of LLM services already, creating a new tier of cloud provider. Emerging evidence for effectiveness of dataset poisoning and existence of embedded, systemic vulnerabilities in models.
Potential impact	**Very high** Potential risk of large blackout events, cyber-physical damage, data breaches or market failures increases significantly as LLM tooling integrates across every sector.

As general-purpose tools which have strong 'comprehension' of human generated content across modalities, it is reasonable to expect LLM integration and impact to occur across many distinct levels of society, with the size of units effected (political, organisational, economic, and cultural) growing proportionately with the capabilities of the tools. Threat actors will potentially be able to attack this new layer of scaffolding in an already densely connected world, creating new opportunities for large accumulations or catastrophes. Widespread integration brings with it a multiplicity of systemic coupling risks, both direct and indirect:

- Provider concentration risk, with the emergence of several monopolistic providers of (legal) LLM tooling for individuals and enterprises (OpenAI, Google, Microsoft) acting as additional single-points-of-failure[10]

- Common source datasets for LLM training create risk of vulnerabilities embedded in models accidentally or intentionally via dataset poisoning.
  - Potential for AI generated code with common vulnerabilities mass generated by coding-assistance tools
  - Potential for wide-spread aberrant behaviour of LLMs triggered by innocuous inputs and services utilising them[2]
- Use of LLM-derived algorithms to control large centralised systems, such as industrial or financial systems, could result in unpredictable failure modes with large footprints of exposure[3]
- Model bias stemming from fundamental inductive biases in algorithm architecture or data may result in systemically correlated decisions, processes, or output across industries where discovery of these correlations is difficult or impossible[4]
- Potential geopolitical tensions arising from control of semiconductor manufacturing capacity, advanced research, or skilled workers amplify uncertainty

[2] Example: It was discovered early in 2023 that specific nonsense strings could cause extreme, unpredictable and sometimes harmful output from GPT3/3.5 models, including through the ChatGPT interface. This has since been patched.

[3] Research utilising LLM applications for control of industrial HVAC systems or electrical grid load balancing has already been published [12]

[4] Example: Firms separately using LLMs for algorithmic underwriting may inadvertently correlate themselves due to inherent architectural inductive bias in how LLMs evaluate evidence or perform analysis.

Considerations for business and insurance

The available evidence, as discussed in the previous sections, allows us to describe the likely impact of Artificial Intelligence and Large Language Models on the frequency and severity of cyber-related losses, providing a strong basis for businesses and the insurance industry to carefully assess the potential impacts.

A new threat landscape

Overall, AI has the potential to act as an augmentation of threat actor capability, enhancing the effectiveness of skilled actors, improving the attractiveness of the unit cost economics, and lowering the barrier to entry. It is likely to mean that there will be more vulnerabilities available for threat actors to exploit, and that it will be easier for them to scout targets, construct campaigns, finetune elements of the attacks, obscure their methods and fingerprint, exfiltrate funds or data, and avoid attributability. All these factors point to an increase in lower-level cyber losses, mitigated only by the degree to which the security industry can act as a counterbalance.

- Initial access vectors which rely on human targets making errors of judgement (spear phishing, executive impersonation, poisoned watering holes, etc) are likely to become significantly more effective as attacks become more targeted and finetuned for recipients
- Attacks are likely to reach broader audiences due to lower cost of target selection and campaign design, meaning the absolute number of losses, and the potential severity of each loss could grow
- Industrial or operational technology attacks are likely to become more common as automation uncovers vulnerabilities
- Embedding AI into software could create entirely new initial access vectors for threat actors to exploit, resulting in larger surface area of attack, and consequently more claims
- The industrialised production of synthetic media content (deepfakes) poses significant challenges for executive impersonation, extortion, and liability risks

Though more companies will be vulnerable to cyber attacks and there will be more security flaws that threat actors can exploit, it is uncertain if this will lead to an increase in highly targeted attacks on specific companies, an increase in broad attacks aimed at many companies, or some other mixed outcome. The increased number of potential targets and vulnerabilities creates the potential for growth in both focused and widespread cyber campaigns.

Overall, it is likely that the frequency, severity, and diversity of smaller scale cyber losses will grow over the next 12-24 months, followed by a plateauing as security and defensive technologies catch up to counterbalance.

Cyber catastrophes

Cyber campaigns tend to be designed with specific objectives and aim to maximise returns for the perpetrators, so most threat actors have a strong incentive to keep their actions concealed and their attacks contained. Catastrophes in cyber occur, for the most part, because the mechanisms put in place by the perpetrators to keep the campaign under control have failed.

The exception to this is state-backed, hostile cyber activity, which includes campaigns designed to cause indiscriminate harm and destruction. It is important to look at the two types of events separately, and distinguish between manageable cyber catastrophes and state-backed, hostile cyber activity.

There is evidence to suggest that the AI-enhancement of threat actor capabilities detailed in the previous section could increase the frequency of manageable cyber catastrophes. However as the mechanism of action is indirect, the magnitude of any increase is likely to be small.

There are several factors driving the occurrence of manageable cyber catastrophes considering AI augmentation of threat actor capabilities:

- The frequency of manageable cyber catastrophes may increase as campaigns are designed to target a broader set of business, coupled with some automation of attacks
- AI-enhancements are also likely to result in better and more effective designs of controls for cyber campaigns. This would allow threat actors to develop more targeted campaigns, meaning that the overall increase in frequency for catastrophes is likely to be lower than the increase for smaller scale losses
- There is evidence of concentration of LLM services, creating a new tier of cloud provider. This new breed of cloud providers would in itself be vulnerable to failures, therefore increasing the frequency of catastrophes associated to Single Point of Failure

The last point deserves some more context. The emergence of LLM services creates an opportunity for threat actors to monetise their attacks in novel ways. For instance:

- Steal all the data that users are submitting to the model endpoint
- Modify the model invisibly to produce biased results in a way that benefits the threat actor (or harm others in ways the threat actor can exploit)
- Steal the models
- Steal all the training data

A concentration of LLM services in turn creates fertile ground for large accumulations, or in other words catastrophes, akin to existing service provider failure scenarios, but with potentially slightly different, and more severe, effects than is possible today.

In conclusion, it is highly probable that the frequency of manageable cyber catastrophes will moderately increase. The risk is very unlikely to sharply escalate without massive improvements in AI effectiveness, which current industry oversight and governance make improbable; this is an area where an increased focus from regulators may be helpful. The increases in catastrophe risk will more likely be gradual based on the steady but incremental progress in AI capabilities that can reasonably be anticipated.

State-backed, hostile cyber activity

State-backed, hostile cyber activity, which includes campaigns designed to cause indiscriminate harm and destruction, gives rise to systemic risks which require a different pricing and aggregation approach.

The effects of Gen AI on this type of systemic risk will surface in the augmentation of tooling and the automation of vulnerability discovery, both of which could enhance existing means to intentionally cause harm and destruction. It is conceivable that the efforts to discover new exploits could concentrate on high impact targets, particularly industrial technology.

The conclusion is that cyber weapons are likely to become more effective, in both destructive power and espionage capabilities. However, it is unclear to what extent the proliferation of advanced capabilities will increase the risk of a major catastrophe happening. The trend is clearly upwards, but once again the human factor will come into play, and the mere existence of these capabilities might not directly translate into deployment, let alone indiscriminate deployment.

Responding to the new risk landscape

- **Insurance:** The potential for a broader set of businesses to be subjected to attacks places a greater emphasis on closing protection gaps for currently underserved audiences, such as SMEs
- **Business:** It will be increasingly important for businesses to invest in the mapping of their critical functions and open-up conversations around cyber defence and restoration capabilities and business continuity planning beyond the risk and information security functions. Organisations using Gen AI should also be able to detail the procedures around how they use it and rely upon it, to evidence with transparency any potential operational risks
- **Government:** Cross-industry working groups around cyber security could offer a platform to share information and learnings in a trusted way. Likewise, collaborating on supply chain failovers could be helpful to reduce the disruption to the economy if a business is impacted
- **Society:** Educating society on cyber hygiene practices, such as zero trust or multi factor authorisation, can reduce susceptibility to social engineering. In addition, education programmes for young people could help to instil good practice and foster the right mindset in the next generation

Taking action

Cyber is one of the most complex and critical risks threatening national security and businesses today, and the dynamics of the risk landscape already poses many challenges. The emergence of AI, particularly advanced generative models, is set to amplify these difficulties as it augments threat actor capabilities and increases systemic risk through concentration of critical technology providers and the broad scope of the services they provide.

Cyber insurance has a key role to play in helping businesses and broader society understand and manage this ever-evolving threat. Tackling it requires continued action, collaboration, and an agile approach. As timescales for action are short, it is important that businesses and the insurance industry take proactive steps to manage the potential consequences of an increasingly uncertain and risky threat landscape.

At Lloyd's, we see significant opportunities for AI to make life easier for our customers and those using our market, but we're also giving thought to the risks AI poses and how we can underwrite them. We will continue to work with governments, regulators, risk experts and insurers to understand and underwrite those risks in an intelligent and sustainable way:

Educating our community

Build awareness and education through Lloyd's Futureset as the growth of AI technology transforms the risk landscape

Partnering with industry

Encourage greater collaboration with governments, regulators and technology companies to better manage AI risks

Engaging policymakers

Work with insurers, governments and others to inform the development of intelligent policy guiderails that can support this important technology

Supporting sustainable innovation

Enable new product development through the Lloyd's Lab that will serve as a springboard to develop new solutions responding to the changing cyber threat landscape

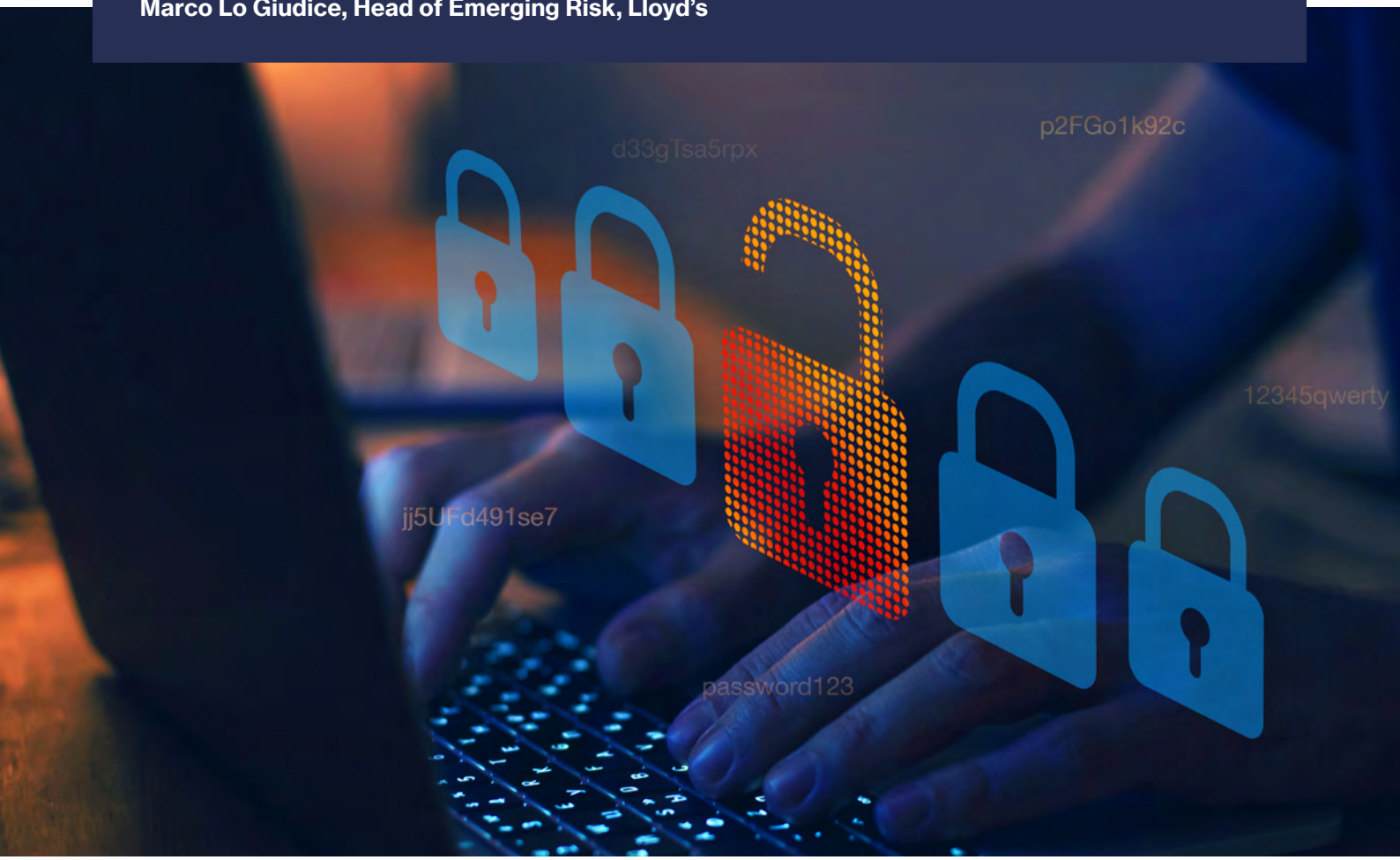
Educating our community

Lloyd's Futureset is our action leadership platform to create and share risk insight, expertise, and solutions to the world's most challenging problems. We will continue to provide insights and educational programmes through Futureset that build greater awareness and understanding as the growth of AI technology transforms the cyber - and the broader - risk landscape.

Case study: Educating our community

In November 2023 Lloyd's, WTW and TWIN (The World Innovation Network) hosted a future-focussed panel discussion between leaders from the worlds of risk, people, computer science and new AI ventures, "AI: utopia, dystopia, or both? A perspective for risk and people". The event explored potential opportunities and threats for the insurance and people business in the evolving AI landscape, and the role of regulation and government to ensure secure and ethical AI adoption. The panel concluded a full day AI event hosted by Lloyd's, attended by over 30 leading AI experts ranging from those in academia, banking, insurance and technology.

> **"Lloyd's is a hub for convening important conversations around future risks, and the opportunities for insurance to help mitigate the associated challenges. It is possible to build AI sustainably using models and data in the correct way - which means we have a big opportunity as a market to face into the challenge of AI innovation. Lloyd's looks forward to continuing to offer its expertise as the technology advances.**
>
> **Marco Lo Giudice, Head of Emerging Risk, Lloyd's**



Partnering with industry

Lloyd's brings together experts who share intelligence, judgement, capital and risk to create a braver, more resilient world. There is a compelling need to gather the experts and trusted actors from across government, industry, capital providers, security agencies and the insurance market to address the current and future threats from cyber. Lloyd's will continue to host events convening these key stakeholders to collaborate, educate, and intelligently create awareness of the developing AI risks associated with cyber.

Case study: Supporting the CISO community

Lloyd's has partnered with the Security Awareness Special Interest Group (SASIG), a networking forum for CISOs and other decisions makers and influencers responsible for security in their organisations, to hold a Risk and Insurance Summit at Lloyd's in April 2024. The Summit convenes the CISO community, insurance industry, cyber security vendors, and public sector stakeholders to highlight the important role of CISOs and cyber security partners in building greater business resilience.

> "The cyber risk landscape is constantly evolving. To appropriately build resilience against its threats, partnerships between insurers, industry, governments and special interest groups like our own are key. We are excited to partner with Lloyd's on a Risk and Insurance Summit in April 2024, to bring together leading experts, share knowledge and encourage dialogue across cyber security and risk mitigation.
>
> **Tarquin Follis OBE, Vice Chairman, SASIG Events**

Engaging policymakers

Lloyd's is committed to working with governments and others to develop intelligent policy guiderails that can support the development of this important technology and create more resilient societies. As part of this we will proactively promote successful applications of AI in insurance and present policymakers and regulators with evidence that supports a pro-innovation regulatory framework for AI. In addition, we will continue to engage policymakers in the UK and internationally to support considered and consistent development of future regulatory frameworks.

Encouraging sustainable innovation

Innovation has been at the heart of the Lloyd's market for over three centuries – writing some of the world's first motor, aviation, satellite and cyber policies, and also the world's first autonomous vehicle insurance. We're all early on in exploring the world of AI, its risks and opportunities – but our market is designed to bring expertise together to underwrite the unknown.

Case study: AI innovation in practice

A member of cohort 10 of the Lloyd's Lab Accelerator programme, Armilla AI focuses on AI assurance and risk management. Its capabilities include evaluating AI model performance, AI audits and due diligence. Armilla AI currently works with AI vendors and enterprises to audit and guarantee the efficacy of their AI products and mitigate risks, and is looking to become an MGA to offer insurance products covering AI risks. Armilla AI drew on the industry knowledge to identify a product opportunity that fit their services, designing two risk transfer products for AI risks during the their time in Lloyd's Lab, AI Product Warranty and AI Liability Insurance, and were able to secure further capacity from Lloyd's syndicates.

The Lloyd's Lab is the at heart of innovation and, as we face the challenges of the AI age, provides a platform to bring together cutting edge InsurTechs, start-ups and ideas which support our market's shared goal of sharing risk to create a braver world. Lloyd's Lab can serve as a springboard for new solutions responding to the changing threat landscape and evolving nature of operational and cyber risk, by:

- Continuing to promote usage of ICX innovation risk code, which is designed to encourage the market to undertake innovation experiments
- Incorporating AI into the themes for future Lloyd's Lab programmes and running a series of pitches with Lloyd's Product Launchpad with a focus on AI-associated losses and risks
- Leveraging Lloyd's Lab network to create a knowledge base of best-in-class companies with solutions, technologies or capabilities that can be utilised by the insurance market to better understand, assess and quantify AI risks

Case study: Fostering sustainable innovation

Futureminds is a Lloyd's Lab innovation programme, designed as a product development bootcamp, looking to power product innovation around significant trends and train a new generation of industry experts in the skills of transforming new ideas into customer-facing products. Venture 6 of Lloyd's Lab Futureminds programme focused on developing insurance products covering the risks associated with applications of AI. The programme saw the development of five product concepts by five cross-industry teams with one product likely to be launched by the market shortly. Lloyd's Lab continue to support the ongoing development of the other propositions, with potential further launches to follow.

> **The Futureminds programme is a true testament that the innovation journey starts with just an idea. Insurance products specifically designed for covering AI risks are still in their nascent stage but as the concepts developed through our AI Venture show, it is clear that insurance will play a crucial role in how this powerful technology is implemented and regulated.**
>
> **Iryna Chekanava, Innovation Growth and Partnerships, Lloyd's Lab**

References

0. Vaswani et al. (2017). Attention Is All You Need. Retrieved from https://arxiv.org/abs/1706.03762

1. Penningtons Law. (2023). The state of AI regulation in the UK and EU. Retrieved from https://www.penningtonslaw.com/news-publications/latest-news/2023/the-state-of-ai-regulation-in-the-uk-and-eu

2. Carnegie Endowment. (2023). Reconciling the U.S. approach to AI. Retrieved from https://carnegieendowment.org/2023/05/03/reconciling-u.s.-approach-to-ai-pub-89674

3. TIME. (2023). OpenAI EU lobbying AI act. Retrieved from https://time.com/6288245/openai-eu-lobbying-ai-act/

4. HM Government. (2023). Industry and national security heavyweights to power UK's frontier AI taskforce. Retrieved from https://www.gov.uk/government/news/industry-and-national-security-heavyweights-to-power-uks-frontier-ai-taskforce

5. Carnegie Endowment. (2023). How hype over AI superintelligence could lead policy astray. Retrieved from https://carnegieendowment.org/2023/09/14/how-hype-over-ai-superintelligence-could-lead-policy-astray-pub-90564

6. Facebook Research. (n.d.). LLAMA model card. Retrieved from https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md

7. OpenAI. (2023). GPT-4. Retrieved from https://cdn.openai.com/papers/gpt-4.pdf

8. Meta. (2023). Llama2. Retrieved from https://arxiv.org/pdf/2307.09288.pdf

9. NSA, FBI, CISA. (2023). Contextualizing Deepfake Threats to Organizations. Retrieved from https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

10. Security Week. (2023). Critical TorchServe Flaws Could Expose AI Infrastructure of Major Companies. Retrieved from https://www.securityweek.com/critical-torchserve-flaws-could-expose-ai-infrastructure-of-major-companies/

11. Fu, Yujia, et al. (2023). Security Weaknesses of Copilot Generated Code in GitHub. Retrieved from https://arxiv.org/abs/2310.02059

12. Song, Lei, et al. (2023). Pre-trained large language models for industrial control. Retrieved from https://arxiv.org/abs/2308.03028

LLOYD'S

Twitter @LloydsofLondon
LinkedIn lloyds.com/linkedin
Facebook lloyds.com/facebook

Lloyd's Futureset