
GuyCarpenter
CyberCube
OUTLOOK ON
AI-DRIVEN
SYSTEMIC RISKS
AND OPPORTUNITIES

Part 1: Introduction

Generative artificial intelligence (Gen AI) is no longer a futuristic concept. It is here, transforming systems to generate unique content, including text, images and even music. Gen AI is undoubtedly transforming the way we work.

There are many ways organizations could revolutionize their respective industries by applying Gen AI to routine business functions. For example, in insurance, Gen AI can assist underwriters evaluating risks by analyzing vast amounts of data, including historical claims, customer information and internal/external cybersecurity factors. By summarizing risk profiles, Gen AI can help underwriters develop the appropriate coverages and make more informed decisions quickly.[1] However, artificial intelligence (AI) technology also presents new cybersecurity risks. While Gen AI can be used to improve operational efficiency, it also opens doors for malicious actors to exploit its capabilities for cyber attacks.

In the August 2024 research paper, Artificial Intelligence: A Multi-Pronged Driver of Cyber Aggregation Risk, co-authored by Guy Carpenter's Cyber Center of Excellence and Marsh McLennan's Cyber Risk Intelligence Center, we discuss 4 new dynamics by which AI deployment can lead to cyber aggregation risk:

1) AI as a software supply-chain threat. Organizations that deploy AI may seek third-party solutions such as ChatGPT, in which the compromise of the vendor model can become a single point of failure for all customers using the model.

2) AI presents a new attack surface. Once AI is deployed, users can interact with the model. Whether it is a chatbot, a claims processing tool or a customized image analysis model, the model receives input and sends outputs. This process is subject to malicious and sometimes accidental manipulation.

3) AI presents a data privacy threat. A model is only as good as the data on which it is trained. To train these models, they must be given access to relevant datasets–often large, sensitive datasets. A compromise to the centralized storage for these datasets can have dramatic downstream effects.

4) AI in security roles. One of the highly touted use cases for AI is in cyber security operations, the type of procedures that require high-level privileges, such as those present in CrowdStrike's recent faulty software update. With such critical response decisions given to AI, the potential for errors or misconfigurations may increase, resulting in additional risks.

While that paper explores these risks from a conceptual, forward-looking perspective, this paper serves as a complement, focusing on the evolving technical and analytical aspects of AI impacts.

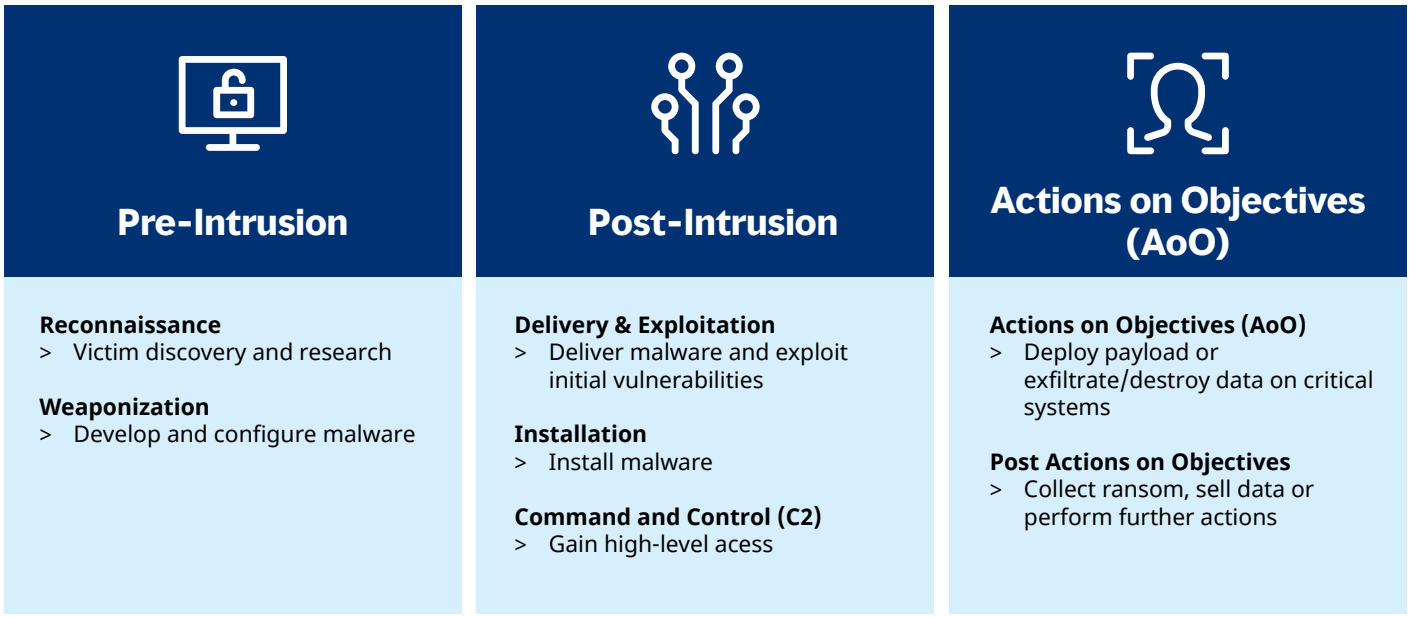
Recognizing the potential exposure accumulation risk arising from AI, it is important for the (re)insurance industry to look ahead and forge an analytical pathway to measure the risk, while embracing the positive side of AI. Partnering with leading cyber risk modeling vendor CyberCube, our study discusses a framework for systemic risk quantification, then investigates 2 counterfactual examples as blueprints for an AI-empowered cyber attack.

Part 2: Exploring an Analytical Framework for AI Risk Quantification

Implications for Cyber Catastrophe Events

To understand the implications AI technologies may have on cyber catastrophe events, we will examine the effects AI will likely have on CyberCube's primary catastrophic model components through the lens of the kill chain. Using the kill chain steps, areas of research are then aligned to model components and attack areas where AI may influence future cyber catastrophe events. Initial research has shown examples throughout the kill chain in which AI could be used in attacks. For the purposes of this paper, we will focus on the areas with demonstrated proof of concept to highlight their relevance.[2]

Figure 1: Kill Chain



Source: CyberCube

Using the Kill Chain

The Kill Chain Approach, popularized by Lockheed Martin,[3] helps to characterize the possible steps in an attack, allowing for decomposition of attack characteristics into a common framework. For the purposes of this exercise, the kill chain framework will allow us to identify where the impact of AI could most likely be felt in the cyber threat landscape. The kill chain consists of 7 stages, with CyberCube adding the eighth stage of Post Actions on Objectives (Post-AoO). These stages are broken into 3 groups, pre-intrusion ("pre"), post-intrusion ("post") and Actions on Objectives (AoO), as shown in Figure 1.

Core Model Areas

The 3 core CyberCube model components to analyze any event are:

• Frequency—How often does the event occur?

• Footprint—Which companies are connected to and impacted by an event?

• Severity—What is the financial loss for impacted companies?

Together, these model components generate an event set, establish which companies are impacted and generate financial loss for each affected company according to cost components for each event in the event set.

Frequency and Footprint Impacts

The frequency and footprint of an event can be detailed in both the pre-intrusion and post-intrusion portions of the kill chain. Early research has shown that an increase in the likelihood of large-scale events could be driven by an increase in the speed and capability of threat actors utilizing AI.

Tools like large language models (LLMs) have been demonstrated to allow for higher-quality social engineering at scale (phishing, deep fakes, etc.), quicker identification of vulnerabilities and the possibility of a larger initial footprint as a result.[4] This may result in attacks reaching scale faster, meaning global catastrophic event frequency could increase overall through a greater number of smaller events being able to reach materiality. Events already considered material enough to be modeled may increase their footprint via the same methods, increasing the average number of companies impacted by events.

This year, proof of concepts and initial threat intelligence reports from threat intelligence company Recorded Future show LLM usage in phishing and social engineering increases the efficiency and efficacy of the reconnaissance, weaponization and delivery stages of attack.[5] Research has also led to the discovery of prompt injection and manipulation of LLMs by adversaries to gain access or perform post-intrusion-type actions on networks

deploying customer-facing LLMs.[6] This shows exposure is not just from attackers that leverage LLMs for their own purposes, but the risk in LLM compromises resulting in "insider threat" events for companies that deploy customer-facing LLMs.

Frequency Impacts

The proliferation of Gen AI as a tool to enhance various attack vectors will certainly increase the frequency of attacks. AI includes the inherent advantage of automation, if it is not already incorporated in the attack. But AI also has the potential to evolve and adapt during attacks as it learns from the experience of previous attempts.

These are distinct advantages compared to traditional attack tools. Whether these characteristics result in net greater frequency of successful attacks will depend on how successful defenders are in developing and deploying defenses based on traditional strategies and defensive AI approaches.

All else being equal, we would expect defenders to have a distinct advantage over threat actors, primarily since legitimate developers of defensive tools will have greater access to superior AI technology and training data from user systems.

However, not all defenders will have the resources or inclination to avail themselves of this advanced technology and, as such, logic dictates that there should be some net increase in the frequency of successful attacks on these less-resourced or less-prepared organizations. The influence of any trend will be difficult to predict, as there are many variables to consider, especially because developments in the AI field are dynamic and uncertain. The most likely outcome is that larger, more resourced or more prepared firms have a better chance at reducing their (often outsized) exposure to cyber risk by deploying AI in defensive mechanisms, while smaller, less-resourced or less-prepared firms will likely have increased exposure to these novel attack trends and methods. This also likely increases the variation of possible impacts from one organization to another when other factors, such as size and industry, are the same.

Cyber threat landscape data suggests that trends in event frequency are often "up and down," meaning increases in event frequency are often followed by relative decreases in frequency. This "wave form" trend is often due to novel attack methods and techniques being countered with advances in defensive methods and capabilities.

An example of this trend is the rise in ransomware that was combated with the closure of certain ports and advances in backup requirements and standards. This trend is anticipated to continue with the advancement of artificial intelligence, but with shortened time between wave peaks as attackers and defenders learn and adapt to one another at faster rates. These effects could counteract one another, especially as the technology matures. Moore's Law (the principle that the speed and capability of computers can be expected to double every two years), and the overall acceleration of semi-conductor advancement, is an example of physical restrictions presenting themselves for transistor technology. The advancement of Gen AI technology will likely be similar. The ability to innovate and improve will likely slow at certain points, leading to periods of reduced offensive and defensive advancement with relative calm.

Frequency Prediction: Potential for greater variation between defender entities that leverage AI versus those that do not.

Efficacy Prediction: Defense versus Offense—Increased volatility.

Footprint Impacts

As discussed above, AI enhancements to attack vectors will increase the efficacy and efficiency of attacks in the pre-intrusion phases of the cyber kill chain. Threat actors will be able to attack a larger number of targets in a more cost-efficient manner, with an expected increase in success rate (through greater targeting of weaker organizations), resulting in a larger footprint for a given cyber threat campaign.

Moreover, we also expect AI to have a significant impact on the effectiveness of the post-intrusion phases of the cyber kill chain. AI can be expected to enhance threat actor capability in target enumeration, lateral movement, privilege escalation and efforts to evade intrusion detection. These enhancements to post-intrusion phases will likely allow attackers to compromise more assets at a greater infection rate, resulting in more significant damage potential. This could lead to more assets being compromised, a larger volume of records exposed in the case of a data breach attack and greater overall leverage in extortion negotiations.

AI ENHANCEMENTS TO ATTACK VECTORS WILL INCREASE THE EFFICACY AND EFFICIENCY OF ATTACKS IN THE PRE-INTRUSION PHASES OF THE CYBER KILL CHAIN

Similar efficiencies can be expected in the action on objectives (AoO) phase, where more efficient data exfiltration processes or data encryption should aid threat actors. AI will also have the effect of causing many post-intrusion and AoO activities to be more difficult to detect by conventional detection and response tools, which will increase an attack's dwell time (the amount of time an attacker is in the system before being detected and removed). Dwell time may be the most significant determinant of the magnitude of impact of many cyber attacks, including data breach and ransomware attacks. As with the pre-intrusion phases, these dynamics may be countered by AI enhancements incorporated in cyber defenses. These developments are complex and difficult to forecast, and many variables and causal relationships are unclear at the moment.

Footprint Impact Prediction: Greater variation between defenders leveraging AI and those that do not.

Malware

Exploitation, command and control, and actions on objectives have been demonstrated through the creation of novel malware with the help of LLMs to exploit known vulnerabilities more efficiently. These changes could affect all types of cyber attacks, from outages to data breaches, to malware and ransomware.

Furthermore, a more nuanced usage of AI in cyber attacks is the mutation of malware through the kill chain—known as polymorphic malware—to avoid common defensive technologies that use pattern recognition or heuristics techniques. While not entirely new, research dating back to 2019 shows polymorphic malware proof of concepts, which can improve their own ability to rewrite themselves to avoid heuristic-based, anti-malware using LLMs.[7] This capability could be used to perform malicious operations at scale using homegrown LLMs. Threat actors could automate the mutation of a virus, allowing it to:

- Increase dwell time, which could lead to greater severity.
- Mutate often enough to avoid signature detection.
- Automate the learning and command and control (C2) processes to spread faster, both externally and internally within networks.

Such changes could improve upon current mutation algorithms in malware that are not as dynamic. The LLMs at the center of the process could then learn just as the defenses themselves learn to attempt to stay ahead. Lateral movement and infection propagation capability would be particularly applicable to ransomware campaigns attempting to extort wider footprints of systems for higher profits. The automation of the command and control on victim networks and the scale of payment and negotiations could be expedited using LLMs.

A proof of concept dubbed "BlackMamba"[8] has already been developed by HYAS Labs using an LLM to "synthesize a polymorphic keylogger functionality on the fly."[9] These examples point to malware that is auto-generated and polymorphic in nature, which will likely lead to increases in overall effectiveness and propagation over time.

Defensive use of Gen AI by cybersecurity vendors and threat intelligence services will also lead to a greater ability for defenders to differentiate malware from normal system operations and identify potentially malicious activity with higher speed and accuracy. These advancements have been widely publicized and quantified, but continually testing these capabilities against real-world attack campaigns will be vital in understanding their efficacy.

AI Impact on Malware: Increased dwell time potential

Data Breaches

Mass exfiltration of data has often been a challenge for threat actors. The ability to extract or exfiltrate large volumes of data at high rates for extortion and sale has often been a barrier to scaling attack profitability. Research has shown machine learning can allow faster and more stealthy data exfiltration by reducing extraction file sizes and automating mass data analysis to identify valuable information within a sea of worthless data.[10, 11] This could result in more effective breaches, which identify valuable crown jewels faster, and extract only valuable data much faster, leading to larger ransoms and increased legal liability frequency and severity.

Defensive AI has also been shown to increase detection of exfiltration activities and unauthorized data access at scale. Continued development of behavioral recognition systems and packet inspection at scale (Zero Trust vendors that require all users to be continuously authenticated and verified before granted access) has been in development for years and will be vital in combating offensive advancements.

AI impact on Data Breaches: Increased exfiltration and monitoring ability

Part 3: Examining AI Implications on Historical Events

Having examined the theoretical ways in which AI can alter the frequency, footprint and impact of a cyber attack, we now investigate 2 counterfactual examples as blueprints for an AI-empowered cyber attack. These examples will concentrate on the application of AI within malware and data breaches, as introduced in Part 2.

Counterfactual 1: Ryuk Ransomware

From 2018 to 2019, Ryuk was a type of ransomware used in many campaigns[12] targeting large, public entities with the goal of financial gain through encryption and ransom payments. During that period, Ryuk accounted for 3 of the top 10 largest ransom demands: USD 5.3 million, 9.9 million and 12.5 million.[13]

Ryuk spread via very targeted means, which included using tailored spear-phishing emails and exploiting compromised credentials to remotely access systems via the Remote Desktop Protocol (RDP). The delivery method for Ryuk was through spam emails, often sent through spoofed addresses, to avoid raising suspicion. Emotet malware, a banking Trojan Horse, was typically used in combination with Ryuk. With RDP, a cybercriminal could install and execute Ryuk directly on the target machine or leverage their access to reach and infect other, more valuable systems on the network. The Emotet loader contained a lot of benign code as part of its evasion techniques and could manipulate security systems to avoid security detection.

With machine learning capabilities,[14] a polymorphic malware can be designed to recursively generate new code variants without human intervention as it calls out to a Gen AI model such as ChatGPT or some more purpose-built utility. The malware itself can periodically create an evolved version of its own malicious code that is more evasive and difficult to detect, utilizing techniques that security tools often are not equipped to handle.

This has the potential to amplify the duration of the infection and its resulting damage. However, malware that has to call out to an external Gen AI model for code updates may be more detectable by security operations teams. A logical progression of such an attack strategy might include a variation of the Living Off the Land (LOTL) technique, where the malware utilizes an internal Gen AI model for polymorphic activity.

For this reason, defenders should secure internal Gen AI models and the data they are trained on, particularly any model used in cyber defense operations. In addition, AI can make it easier for perpetrators to design new malware variants. Instead of taking months to upgrade the malware, they can leverage AI to train models on vast datasets of malware samples to learn patterns and devise new strategies for mutation in a shorter period. These models can then autonomously generate new variants with altered code structures, effectively staying one step ahead of security defenses.

Takeaway: AI could boost the efficiency of malware, with the potential for an increased likelihood of cyber incidents. The implications of AI-driven polymorphic malware are profound and pose a larger systemic potential for the (re) insurance industry if the risk is not carefully mitigated.

Counterfactual 2: Equifax Data Breach

In 2017, at the time of attack, the Equifax breach was the second-largest breach in history, impacting 163 million records worldwide, including almost half of all Americans. This was only overshadowed by the 2016 Yahoo breach, which still holds the top spot as the most impactful data breach event.

However, the completeness and sensitivity of the data exfiltrated from Equifax give extra weight to the severity of the event. Since 2017, there have been several other significant breaches that outdid the Equifax breach in terms of records impacted (Microsoft Exchange, Facebook). The continued occurrence of these events illustrates that data exfiltration attacks continue to be a significant concern, especially when combined with the additional capabilities of AI and LLMs.

Successes at the 2016 DARPA and DEF CON network defense events have proven that AI can be used to scan for unidentified vulnerabilities (2021 Microsoft Exchange

attack) as well as locate known vulnerabilities similar to the unpatched software used to perpetrate the Equifax breach. At the 2024 DEF CON, AI was used to identify and patch vulnerabilities but, alternatively, could also be leveraged to locate and exploit new vulnerabilities across multiple targets or credit bureaus as opposed to the targeted Equifax breach. The inclusion of AI capabilities to find and exploit vulnerabilities would intensify and broaden the impacts of breaches like Equifax by allowing for event scalability.

A key factor in the Equifax breach was the perpetrators' ability to find and utilize unsecured credentials to gain access to 48 databases. LLMs are capable of identifying files containing unsecured credentials with greater speed and accuracy than an unassisted threat actor. During the Equifax breach, the hackers ran 9,000 queries against the databases, of which only 265 came back with Personally Identifiable Information (PII). If AI had been included, it would have searched for PII with greater precision and highlighted likely instances of valuable data to the attackers, significantly increasing the breadth of the data exfiltrated from the databases and proceeding with greater efficiency.

The Equifax attack ended with the renewal of a lapsed Secure Sockets Layer (SSL) certificate. The update allowed Equifax's information security team to view traffic from the Equifax system and identify suspicious activity. The Equifax team became almost immediately aware of suspicious traffic to Chinese and Chinese-operated IP addresses, leading them to shut down the impacted service. AI could be used to mimic legitimate network traffic and avoid any suspicious increases in network activity that could be used as a warning flag by internal security tools. Disguising the outgoing network traffic with legitimate business operations could have easily extended the time in network during the Equifax breach.

Takeaway: The addition of AI tools can significantly increase the effectiveness of a hacking group by encouraging more efficient lateral movement and greatly broaden the impact in terms of the amount and level of sensitive data exfiltrated. The intersection of AI and an environment of persistent software vulnerabilities creates an opportunity for Equifax-type breaches with greater scalability and intensity.

Part 4: Conclusions and Looking Ahead

This report focused exclusively on identifying traditional cyber perils that could be enhanced by the use of AI tools in attack campaigns. Another key aspect of AI impact on insurance portfolio accumulation risk is AI technology itself being a target (i.e., AI as a single point of failure (SPoF)). The consideration of AI as a SPoF is challenging due to the combination of AI's complexity, its inherently unpredictable and evolving nature, the relatively novel nature of its adoption, its dependency on data and the criticality of the AI and the systems it is integrated with to a business's core operations. Understanding the interplay of these factors, among others, will have a strong influence on enabling reliable risk quantification of AI as a SPoF.

This is a complex topic that requires more research and a thoughtful analytical approach. While Gen AI has been incorporated into CyberCube's Attritional Loss Model (ALM) updates, cyber catastrophic modeling is another consideration. In-depth evaluations have been performed and will continue to be performed to evaluate when AI-centered SPoF events will rise to the level of cyber catastrophes and, as a result, be included in cyber catastrophe risk model updates.

While much of the dialogue on artificial intelligence in cybersecurity has focused on the negative ramifications, we would be remiss if we did not mention its positive contributions to the field. Namely, initial research has shown promising advances in malware recognition and containment, data tagging and monitoring and loss prevention as a whole.

The core of any AI system, like a large language or machine learning model, is the data used to train it. On this ground, defenders have the upper hand. They are playing on home turf, training their models on their environments and partnering with vendors to see the whole picture of the threat landscape to create a custom defensive solution.

The attacker can only see what is outward-facing and must infer the rest. As AI is integrated into endpoint and extended detection and response (EDR/XDR) and cloud security platforms, these models are continuously training on the data from defender networks as well as current threat intelligence data. The effects of AI being leveraged in defensive mechanisms must also be reflected in future cyber modeling frameworks, in order to avoid the risk of overstating the potential ramifications of AI threats.

As AI technology becomes increasingly integrated into our lives, the (re)insurance industry has a unique opportunity to assist policyholders preparing for potential threats arising from AI. In this paper, we began exploring an analytical framework to quantify AI-related risks using CyberCube's kill chain methodology. Guy Carpenter and CyberCube are committed to the continued efforts of developing a concrete path toward assessing and quantifying these risks, which we will explore in greater detail in a future research paper.

Authors

Jess Fung, Managing Director and North American Cyber Analytics Lead, Guy Carpenter

MJ Teo, Vice President and Senior Cyber Actuary, Guy Carpenter

Richard McCauley, Vice President and Senior Cyber Catastrophe Advisor, Guy Carpenter

Andrew Kao, Director of Product Marketing, CyberCube

Joshua Knapp, Cyber Risk Modeling Team Lead, CyberCube

Richard DeKorte, Cyber Security Consultant, CyberCube

1. Generative AI in Insurance: Top 5 Use Cases (appian.com)
2. DarkTrace State of AI Cyber Security 2024 (White paper)
3. Lockheed Martin (Cyber Kill Chain)
4. Department of Health and Human Services (HHS) Office of Information Security (Report)
5. Recorded Future examples (1) (2) (3)
6. Recorded Future on Britain's NCSC LLM (warning)
7. A Survey on Artificial Intelligence in Malware as Next-Generation Threats (Troung & Zelinka)
8. BlackMamba Polymorphic LLM based Malware (SentinelOne)
9. AI-Powered "BlackMamba" Keylogging Attack Evades Modern EDR Security (Dark Reading)
10. A Survey on Artificial Intelligence in Malware as Next-Generation Threats (Troung & Zelinka)
11. DarkTrace State of AI Cyber Security 2024 (White paper)
12, 13. RYUK Ransomware (Trend Micro)
14. Harnessing AI for Polymorphic Malware: The Evolution of Cyber Threats | by Ashley Jackson (Medium)

About CyberCube

CyberCube is the leading provider of software-as-a-service cyber risk analytics to quantify cyber risk in financial terms. Driven by data and informed by insight, we harness the power of artificial intelligence to supplement our multi-disciplinary team. Our clients rely on our solutions to make informed decisions about managing and transferring cyber risks. We unpack complex cyber threats into clear, actionable strategies, translating cyber risk into financial impact on businesses, markets, and society as a whole.

CyberCube was established in 2015 within Symantec and now operates as a standalone company. Our models are built on an unparalleled ecosystem of data and validated by extensive model calibration, internally and externally. CyberCube is the leader in cyber risk quantification for the insurance industry, serving over 100 insurance institutions globally. The company's investors include Forgepoint Capital, HSCM Bermuda, and Morgan Stanley Tactical Value. For more information, please visit www.cybcube.com or email info@cybcube.com.

About Guy Carpenter

Guy Carpenter & Company, LLC is a leading global risk and reinsurance specialist with 3,500 professionals in over 60 offices around the world. Guy Carpenter delivers a powerful combination of broking expertise, trusted strategic advisory services and industry-leading analytics to help clients adapt to emerging opportunities and achieve profitable growth. Guy Carpenter is a business of Marsh McLennan (NYSE: MMC), the world's leading professional services firm in the areas of risk, strategy and people. The company's more than 85,000 colleagues advise clients in over 130 countries. With annual revenue of $23 billion, Marsh McLennan helps clients navigate an increasingly dynamic and complex environment through four market-leading businesses including Marsh, Mercer and Oliver Wyman. For more information, visit www.guycarp.com and follow us on LinkedIn and X.