



AI RISK AND THREATS
(米国における AI のセキュリティ脅威・
リスクの認知調査レポート)

2024 年 5 月 独立行政法人情報処理推進機構

Prepared by Next Peak

Table of Contents

I.	Executive Summary.....	3
II.	Purpose, Scope, and Methodology.....	4
III.	Evolution of AI.....	6
A.	Overview.....	6
B.	Increase in AI Systems’ Capabilities and Deployment.....	8
1.	Machine Learning.....	8
2.	Generative AI.....	9
3.	Malicious Models and Jailbreaks.....	12
C.	Summary.....	14
IV.	AI Threats and Risks.....	15
A.	Introduction.....	15
B.	AI-Enhanced Traditional Cyberattacks.....	15
1.	Overview.....	15
2.	AI Enhancements.....	15
3.	Risks and Impacts.....	19
C.	AI-Enabled Disinformation.....	20
1.	Overview.....	20
2.	AI Enhancements: Generation and Dissemination.....	20
3.	Risks and Impacts.....	22
D.	AI-Enabled Disruption or Maloperation of Systems.....	25
1.	Overview.....	25
2.	AI System Threats.....	26
3.	Risks and Impacts.....	27
E.	AI-Enabled National Security Threats.....	28
1.	Overview.....	28
2.	AI Enhancements.....	28
3.	Risks and Impacts.....	34
F.	Business Risks Due to Misuse of Generative AI.....	37
1.	Overview.....	37
2.	AI-Related Threats.....	38
5.	Risks and Impacts.....	42
V.	Analysis and Key Findings.....	43
1.	Analysis of Research.....	43
2.	Key Findings from Expert Interviews.....	45
3.	AI Threat and Risk Chart.....	47

6. Expert Interviews	49
1. Chief Architect 1.....	49
2. Public and Private Sector Expert 1	51
3. Public and Private Sector Expert 2	53
4. National Organization Executive 1	55
5. Financial Sector CSO/CISO 1	56
6. Cyber Security Company CEO 1	58
7. Cyber Security Company CEO 2	60
8. National Security Expert 1.....	62
9. Cyber AI Research Expert 1	65
10. Venture Capital Investor and Philanthropist 1	67
7. Appendix	69
Appendix 1: Table of Recent Examples of LLM-Themed TTPs by APTs	69
Appendix 2: Table of AI-Enabled Disinformation Efforts to Undermine Democracy and/or Increase Censorship.....	71
8. Annotated Bibliography	72
9. References	75
10. Footnotes	80

I. Executive Summary

Artificial Intelligence (AI) technologies are continuously advancing, leading to extensive impacts on individuals, businesses, and society at large. In particular, the widespread adoption of generative AI has ushered in a new era of technological advancements, offering innovative solutions across various sectors. Since its mainstream emergence in late 2022, generative AI tools—including generative text, audio, image, and video—have become critical to industry operations, enhancing organizational efficiency, productivity, and profitability. Research suggests that generative AI has the potential to contribute between \$2.6 trillion to \$4.4 trillion annually to the global economy.¹

At the same time, the democratization of AI poses significant challenges for cybersecurity, exacerbating existing risks and presenting novel threats. This report conducts a survey of over 80 sources and 10 expert interviews to investigate AI risks in five areas: 1) AI-enhanced traditional cyberattacks; 2) AI-enabled disinformation; 3) AI-enabled disruption or misoperation of systems; 4) AI-enabled national security threats and 5) business risks due to misuse of generative AI. The report evaluates the present and emerging applications of AI, the risks they pose, and the potential impacts.

The proliferation of AI tools is poised to intensify current cybersecurity challenges from enhanced cyberattacks to disinformation campaigns. Both state and non-state threat actors are swiftly adapting, leveraging AI to enhance the efficiency and impact of their operations. Notably, generative AI tools have lowered the barrier to entry for cybercriminals, leading to an increase in more sophisticated and personalized attacks: the democratization of AI enables the acceleration and amplification of traditional cyber threats, potentially outpacing defenders' abilities to adapt and respond effectively.

Generative AI tools are reshaping the disinformation landscape, particularly in the form of state-sponsored campaigns for election interference, domestic disinformation, and mass surveillance. As generative AI tools become more accessible and advanced, the proliferation of deepfakes exacerbates the normalization of disinformation, eroding trust in institutions and democratic processes over time. Governments and social media platforms alike face challenges in countering false information while preserving the free flow of information.

As businesses increasingly adopt generative AI tools, they are confronted with a growing risk from biased content generation to potential data poisoning attacks. These risks present significant challenges to the reliability and safety of critical systems and underscore the urgent need to prioritize the safeguarding of AI tools to mitigate the severe consequences of false outputs, inadvertent escalation, and perpetuated biases.

The report concludes with a set of key findings from the research and interviews such as the force-multiplying impact of AI and the effect of AI's democratization. The report also offers an analysis of the AI threats and risks through a chart analyzing the threat, risk, impact, timeline, and key concerns of AI-enabled threats.

II. Purpose, Scope, and Methodology

The proliferation of artificial intelligence (AI) technologies presents opportunities for innovation in business and daily life. Despite the potential benefits, new AI technologies also bring new security risks that can exacerbate existing cybersecurity challenges. In 2021, Next Peak provided the Information Technology Promotion Agency of Japan (IPA) with an overview and analysis of cyber risks relating to the application of AI technology. Since 2021, the evolution and progress of AI development have been exponential, requiring additional research and an updated report. Thus, this report provides an updated survey that analyzes the rapid evolution of AI technologies in the past two years and addresses new and intensifying risks. The report draws from existing research and literature, ten expert interviews, as well as previous analysis.

The report seeks to inform the new Japan AI Safety Institute (AISI Japan) which was established in February 2024. The new institute is aimed at studying evaluation methods for AI safety, designed to be a counterpart of the United States (US) Artificial Intelligence Safety Institute (AISII) at the National Institute of Standards and Technology (NIST).² Due to the extensive and expanding use of AI across various industries and digital domains, this report focuses on specific AI risks in coordination with IPA's priorities. We focus primarily on the following five risks: 1) AI-enhancements of traditional cyberattacks; 2) AI-enabled disinformation; 3) AI-enabled disruption or mis-operation of systems; 4) AI-enabled national security threats; and 5) Business risk due to incorrect use or misuse of generative AI. Within these five areas, we assess how AI intensifies cybersecurity, national security, and business risks by exploring the current and potential uses of AI for adversarial purposes and the vulnerabilities of misuse or maloperation of deployed AI systems. Furthermore, we assess the broader impacts of these risks on society, industry, and individuals.

This report was produced in multiple phases:

1. First, an extensive literature review of the evolution of AI technologies and the cybersecurity risks that they create was conducted. Since the 2021 report, there has been a significant increase in research, publications, and articles about AI-enabled cyber risks. Drawing on previous research and the existing literature, the report considers how the rapid evolution of AI technologies presents new and intensifying threats to our society through malicious intent, erroneous implementation, and lack of oversight or regulations. In the literature review, over 80 individual sources were surveyed. The annotated bibliography highlighting the most significant 10 sources can be found at the end of this report.
2. Next, previous research, current literature review, and existing interviewee base were scoped to identify a list of AI experts to engage. The process informed the development of key interview questions tailored to address these critical issues. Subsequently, interviews with a diverse selection of AI experts from across the public, private, and non-profit sectors were conducted. The interviewees' expertise and perspectives were also wide-ranging, covering technical AI development to legal and policy frameworks to national security. Readers can find individual interviews and short bios of each interviewee in the Expert Interviews section.

3. While the interviews were being conducted, we developed a set of initial key findings and insights from the latest research in our literature review. The next section titled “Evolution of AI” reflects the changes in cyber risk due to AI since the 2021 report for IPA and specifically highlights the changes in AI deployment and use for malicious purposes. The following sections further expand upon specific AI-enabled risks and conduct an analysis based on sector, timeline, and severity, in accordance with the five focus areas outlined in this report:
 - AI-enhanced traditional cyberattacks
 - AI-enabled disinformation
 - AI-enabled disruption or maloperation of systems
 - AI-enabled national security threats
 - Business risk due to incorrect use or misuse of generative AI
4. Lastly, the Analysis and Key Findings section synthesizes the findings from the literature review with insights from expert interviews and presents a chart of AI-enabled threats.

III. Evolution of AI

A. Overview

AI is a general-purpose digital technology that is transforming various aspects of human life, industry, and science. Although definitions vary, in general, AI refers to a broad discipline of creating intelligent machines as opposed to the natural intelligence demonstrated by humans. The landscape of AI has experienced a surge in capabilities and applications over the last few years, leading researchers and companies to eagerly adopt automation and swift decision-making processes. AI models and systems are complex, making categorization of different systems and models difficult. Below is an attempt at mapping the complexity of AI technology and systems.

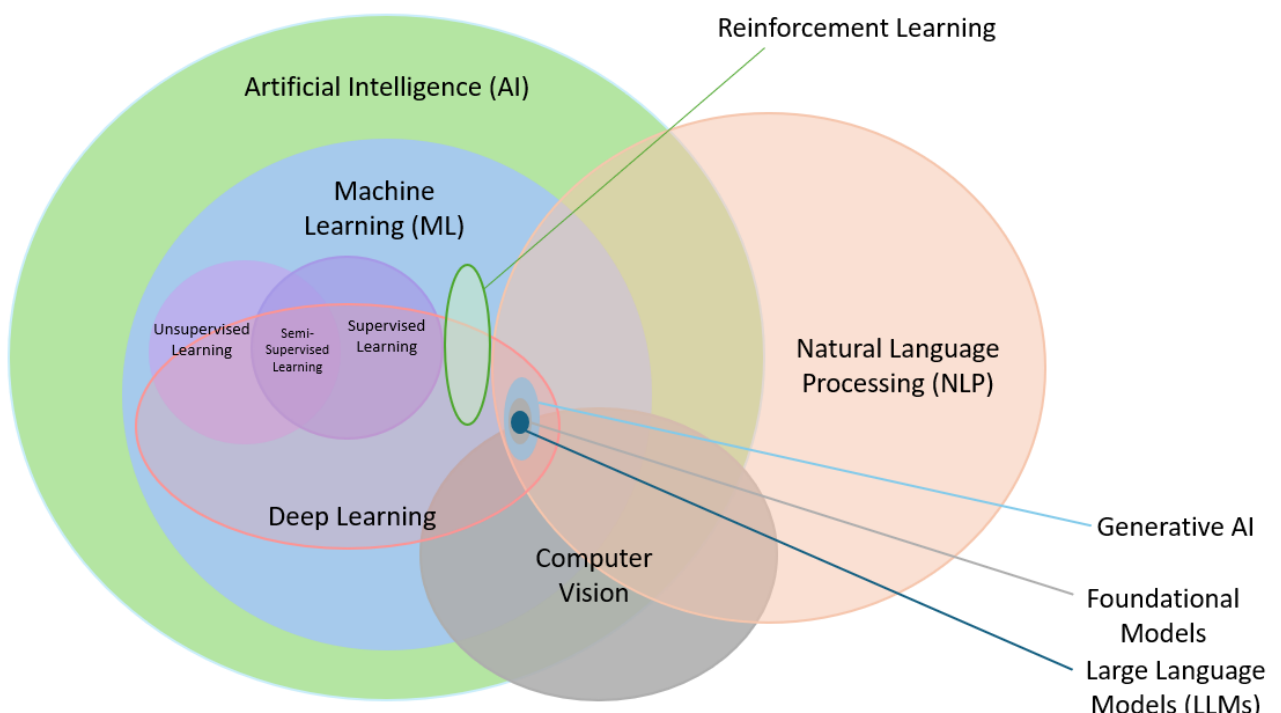


Figure 1: Diagram of AI systems and technologies³

In the past few years, machine learning (ML) technology has made particularly significant progress, with generative AI tools being widely adopted. Within generative AI, Large Language Models (LLMs) have been widely democratized with the launch of Chat GPT.⁴

- **Machine Learning (ML):** a subset of AI that often uses statistical techniques to give machines the ability to "learn" from data without being given explicit instructions. This process involves training a model with a learning algorithm that enhances the model's performance on a specific task.
- **Generative AI:** A family of AI systems that can generate new content based on prompts. Generative AI techniques are widely used in tasks such as image, text, and audio generation.

- **Large Language Model (LLM):** a model trained on vast amounts of, often, textual data to predict the next word in a self-supervised manner. The term “LLM” is used to designate multi-billion parameter language models (LMs), but this is a moving definition. Notable examples of LLMs include:
 - **GPT (Generative Pre-trained Transformer):** The GPT series—such as GPT-2, GPT-3, and GPT-4—are known for text generation capabilities and are commonly used for tasks like language translation, content generation, and chatbots.
 - **BERT (Bidirectional Encoder Representations from Transformers):** BERT models are designed for natural language understanding and perform tasks like sentiment analysis and question-answering.

Thus, this report includes a deep survey of ML and generative AI. Other subsets or specific techniques within the field of AI include:⁵

- **Deep Learning:** an approach to AI inspired by how neurons in the brain recognize complex patterns in data.
- **Foundational Model:** AI systems with broad capabilities that can be adapted to a range of different yet more specific purposes. The original model provides a base or a “foundation” on which other models can be built.⁶ Foundational models are trained on vast datasets and are adaptable to various downstream applications which allow for increasing AI integration across a range of industries and fields.⁷
- **Reinforcement Learning (RL):** the process of training machines through trial and error to teach the model to take the best action by establishing a reward system.
- **Supervised Learning:** the process in which the human-structured or labeled data enables the algorithm to extract features from the data.
- **Unsupervised Learning:** the process of the model making its own prediction tasks such as trying to predict each successive word in a sentence without human-structured or labeled data.
- **Semi-supervised Learning:** the process that combines supervised and unsupervised learning, using both labeled and unlabeled data to train AI models for classification and regression tasks.
- **Natural Language Processing (NLP):** the ability of computer systems to understand text. NLP is used in various AI systems like PaLM, GPT-3, and GLM-130B. These systems are trained on large amounts of data and are adaptable to a wide range of downstream tasks.
- **Computer Vision:** a subfield of AI that teaches machines to understand images and videos. Such technologies have important real-world applications, such as

autonomous vehicles, crowd surveillance, sports analytics, and video game creation.

Finally, with the evolution of AI technology, malicious models that copy the models of legitimate AI tools and jailbreak methods are emerging. This section will conclude with a brief overview of those malicious and jailbroken AI models.

B. Increase in AI Systems' Capabilities and Deployment

1. Machine Learning

Advances in ML algorithms have driven progress in AI with a growing set of available data, improvements in algorithmic approaches, and advancements in computing processing power and data storage. These advancements have improved statistical computing models. In recent years, the scale and cost of LLMs, which are tools for ML, have surged, leading to a significant development in ML as well. GPT-2, introduced in 2019 as one of the first LLMs, had 1.5 billion parameters⁸ and cost nearly \$ 50,000 to train.⁹ In contrast, PaLM, a leading LLM from 2022, featured 540 billion parameters, with a training cost of about \$8 million. PaLM is about 360 times larger and 160 times more costly than GPT-2.¹⁰

As the financial demands for AI projects have risen sharply, ML has transitioned from a domain predominantly influenced by academia to one largely shaped by industry innovation. Until 2014, academic institutions released the most significant ML models, but as of 2022, this trend has reversed: industry produced 32 significant ML models while academia only produced three.¹¹

Number of Significant Machine Learning Systems by Sector, 2002–22

Source: Epoch, 2022 | Chart: 2023 AI Index Report

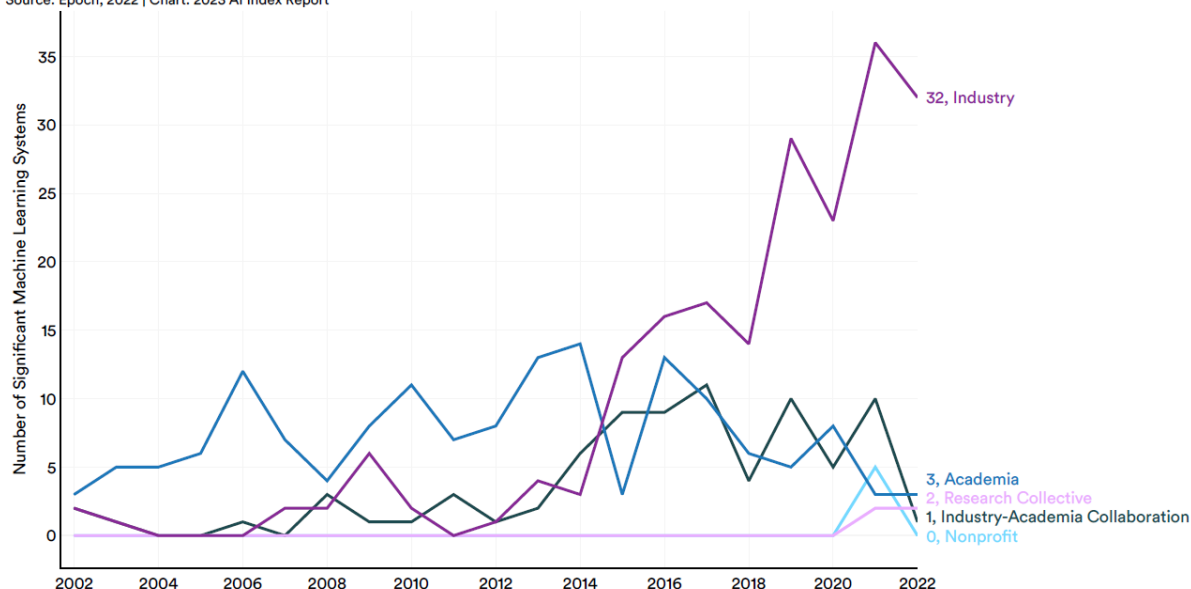


Figure 1.2.2

Figure 2: Significant machine learning systems by sector, 2002-22.¹²

ML innovation is concentrated in the United States (US), with the US creating 16 significant ML systems¹³ in 2022, followed by the UK with eight and China with three.¹⁴ Yet, the US and China have become leading collaborators in AI research, with research quintupling since 2010,

although the pace of collaboration has somewhat slowed in recent years.¹⁵ This trend seems paradoxical as the US and China race for leadership in AI technologies, while researchers increasingly see benefits from sharing expertise.

2. Generative AI

Public use of generative AI has proliferated since late 2022 with the launch of chatbots like ChatGPT by OpenAI, text-to-image systems like DALL-E 2 and Stable Diffusion, and text-to-video systems like Make-a-Video.¹⁶ While ChatGPT’s capabilities were similar to its predecessors such as GPT-3, ChatGPT enabled the everyday user to utilize AI technology, reaching 100 million monthly active users within two months of its launch. New models continue to build on ChatGPT’s success, and just months after ChatGPT was first released, OpenAI released its new LLM, GPT-4, with improved capabilities. Despite these significant advances, generative AI models are susceptible to generating false information, frequently exhibit biases, and can be manipulated to fulfill malicious purposes, underscoring the complex ethical dilemmas linked to their widespread use.

Industry leaders are increasingly integrating generative AI tools into their organizations to enhance organizational efficiency, productivity, and profitability. According to the 2023 McKinsey Global Survey, 79% of respondents reported exposure to generative AI, either professionally or personally.¹⁷ Of these, 22% noted regular use in their work.¹⁸ Notably, 40% of respondents indicated plans to increase their overall AI investment, underscoring the significant advancements in generative AI technology.¹⁹

Respondents across regions, industries, and seniority levels say they are already using generative AI tools.

Reported exposure to generative AI tools, % of respondents

Select demographic:

■ Regularly use for work
 ■ Regularly use for work and outside of work
 ■ Regularly use outside of work
■ Have tried at least once
 ■ No exposure
 ■ Don't know

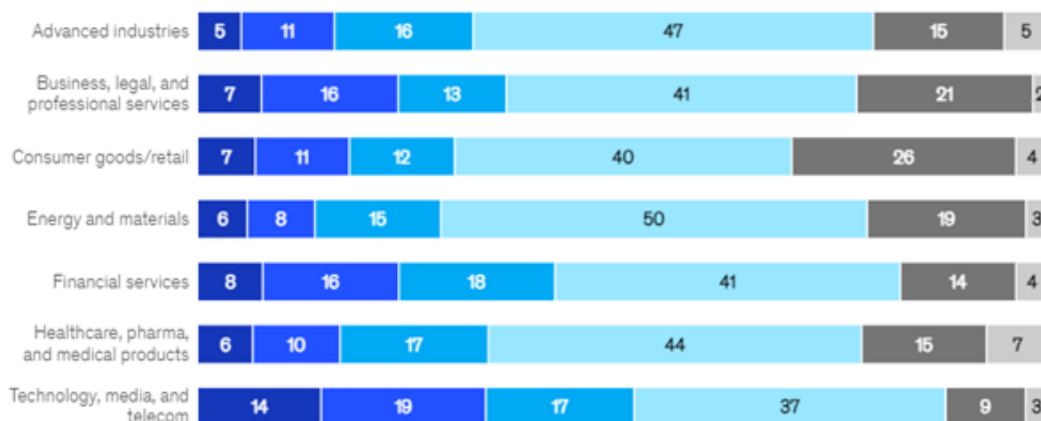


Figure 3: Distribution of respondents using generative AI tools, by industry.²⁰

The advent of generative AI tools, predominantly chatbots, has rapidly altered the threat landscape, with cybercriminals and nation-state actors leveraging these technologies for

malicious activities such as AI-enabled phishing, social engineering attacks, and large-scale disinformation campaigns. To counter these threats, cybersecurity vendors are increasingly integrating AI solutions to detect and mitigate malicious AI usage, bolstering cybersecurity defenses.²¹

i. Generative Audio

Over the past decade, there have been significant advancements in AI-generated audio applications, text-to-speech generation, sound creation, and audio editing. However, until recently, these developments lagged behind those of image and text generation. Recent applications have demonstrated improvements in foundational models and training procedures. These applications have adapted from technology for text-to-image generation, which has significantly improved generative audio quality, controllability, inference speed, and output length. In the last year alone, several text-to-audio models have been released by large tech companies, including Google’s AI test kitchen, Meta’s music generator “Voicebox”, Make-An-Audio by ByteDance, and VALL-E by Microsoft. Underscoring the rapid evolution of generative audio, Microsoft’s VALL-E was trained on 60,000 hours of English speech from over 7,000 speakers, 10 times larger than datasets used by previous text-to-speech systems.²²

Generative audio technologies have had the largest impact on music and film production as well as other creative domains. Advancements in generative audio allow the creation of original music compositions for social media platforms, personalized playlists, royalty-free music, remixes, podcasts, and audiobooks.²³ For instance, Spotify is seeking to use LLMs to replicate non-music content—such as podcasts and audiobooks—and increase its profit margins.²⁴ Additionally, generative audio is increasingly being used to replace audio content in movies, videos, and games.

Generative audio has been reportedly used for adversarial purposes by nation-states, cybercriminals, and extremist groups as part of disinformation campaigns, cyberattacks, or propaganda. Increasing instances of audio fakes have been identified on social platforms like TikTok and YouTube, and platforms are beginning to implement mitigation efforts such as requirements to disclose AI-generated or manipulated media. Notable examples include instances where voice cloning technology, sourced from AI Startup ElevenLabs’ free text-to-speech generator²⁵, has allegedly been used for malicious purposes: a robocall impersonated President Biden to discourage New Hampshire voters from voting in the state’s primary election²⁶, and a TikTok video featured synthetic audio of former President Obama defending a conspiracy theory.²⁷

Affiliates and sympathizers of al-Qaeda and the Islamic State have also adapted quickly to generative audio. For example, the groups are using voice clone technology to create audio deep fake Nasheeds,ⁱ marking an increasingly sophisticated tactic to expand the reach of their content on social platforms such as TikTok. Likewise, users on 4chan—a right-wing message board—reportedly adapted tools from ElevenLabs to generate an audio fake of actor Emma Watson reading an anti-Semitic speech.²⁸

ⁱ Nasheed refers to a song without musical instruments with lyrics that resemble hymns that praise God (Allah).

ii. Generative Video and Photos

Text-to-video generation has advanced rapidly with new generative AI tools to create, edit, and translate videos as well as the ability for face swapping and deepfake visual effects. In late 2022, the first high-quality text-to-video models began to appear, marking an impressive advancement, though still only capable of generating videos of a few seconds duration.²⁹ By February 2024, OpenAI announced Sora—an AI model that can generate realistic and imaginative scenes from text prompts.³⁰

The rise of generative video has also given rise to ethical concerns with the emergence of deepfakes. These manipulated videos can convincingly depict individuals engaging in actions or speech they never actually did. As AI algorithms advance, the risks of disinformation and manipulation also increase. According to a 2022 survey of cybersecurity and incident response professionals, 66% said they had experienced a security incident involving deepfake use in the prior 12 months, marking a 13% increase from the previous year and highlighting the rapid proliferation of such tools.³¹

Likewise, text-to-image generation has also gained widespread attention with the introduction of models such as OpenAI's DALL-E 2, Stability AI's Stable Diffusion, Meta's Make-A-Scene, and Google's Imagen. Generative image technology has advanced substantially to the point where it has become challenging for the average person to differentiate between a real human face and one generated by AI. The entertainment, media, marketing, e-commerce, and sales industries have widely adopted generative images and videos to create unique and personalized content quickly without using specialized equipment, editing expertise, or even actors.

Generative image and video systems raise ethical concerns regarding their tendency for bias along racial and gender dimensions, as well as their potential vulnerability to being jailbroken. For instance, in late January 2024, explicit AI-generated images of Taylor Swift, which allegedly originated on 4chan, were spread across social media. Social media analytics firm Graphika also discovered a series of messages on 4chan urging individuals to circumvent safeguards implemented by generative image tools such as OpenAI's DALL-E, Microsoft Designer, and Bing Image Creator.³² Finally, artists have raised concerns of AI models infringing on their intellectual property: in 2022, three artists formed a class to sue several generative AI platforms, alleging the platforms used their original works without license to train their AI in their styles, allowing users to create works that closely resembled their protected works.³³

There have also been numerous examples of nation-state or state-sponsored actors utilizing generative images and video as part of malign influence operations. In December 2023, the Australian Strategic Policy Institute (ASPI) reported that Mandarin-speaking actors conducted the "Shadow Play" campaign with AI voiceovers across at least 30 YouTube channels to promote pro-China and anti-US narratives.³⁴ Moreover, actors with affiliations to Russia, Iran, and China have engaged in cyber-enabled influence operations, including the creation and dissemination of deepfake videos targeting political figures. Notably, in March 2022, Russia-aligned threat actors defaced several Ukrainian websites with a deepfake video of Ukrainian President Volodymyr Zelensky encouraging Ukrainians to surrender to Russian forces.³⁵

iii. Generative Text

Generative text systems have made significant strides, fueled by the availability of vast data resources and increased computing power.³⁶ Advancements in generative text models are underpinned by neural network language models, including GPT and BERT, which learn how words are used in different contexts by sifting through the patterns in naturally occurring texts. Such models consist of billions of parameters and can process large quantities of data: GPT-3 can process over one trillion words. By arranging probable sequences of words, many of these models can produce text passages that closely resemble human-created writing such as news articles, poems, fiction, and even computer code.

Since late 2022, generative text models have expanded into the public consciousness and have since become essential across different industries including customer service, education, content creation, healthcare, entertainment, and professional services. The latest generative text systems provide a multitude of functions including code generation, software development, product development processes' enhancements, fraud detection, risk management efforts, and synthetic data generation for training and testing purposes. Increasingly, wealth management firms such as Morgan Stanley, Goldman Sachs, BlackRock, and JPMorgan Chase are announcing the integration of ChatGPT-like software to create generative AI assistants for code testing and generation as well as to advise clients on investments.

Alongside these legitimate uses, the use of generative text for malicious purposes has also proliferated. Generative text has become a new avenue for cybercriminals to produce malicious code, personalized phishing emails,ⁱⁱ and generate content for disinformation campaigns at scale. The increasing adoption of generative text models by businesses also creates a growing risk of concentrated data in specialized AI systems, which are seen as prime targets by cybercriminals. These risks are further described in the following parts of this report.

3. Malicious Models and Jailbreaks

With the evolution of AI systems, malicious actors have begun developing malicious AI models as well as jailbreaking existing models, posing a threat to the cybersecurity landscape and enabling cybercriminals to conduct sophisticated cyberattacks at rapid speed and scale. These AI tools lower the barrier to entry for adversaries and hackers offering features like generating malicious code and bespoke malware. While some models have been developed exclusively for offensive cyber purposes, others were initially developed by cybersecurity researchers and intended for dark web research. Allegedly, threat actors have accessed these AI models and have been exploiting them as well.

i. Malicious AI Models

AI experts are already seeing an emergence of malicious AI models. This section summarizes the findings. Various malicious models parallel the modes utilized for GPT.

ⁱⁱ A cyber AI research expert referred to [an industry report](#) during her interview, highlighting that phishing emails increased by 1000% since the introduction of ChatGPT.

WormGPT is a malicious chatbot based on the GPTJ language model, which was developed in 2021 and trained on malware-related data specifically for malicious activities. The model is equipped with a customized LLM that enables cybercriminals to execute various attacks including business email compromise (BEC) attacks.³⁷ This tool presents a range of features including unlimited character support, chat memory retention, and code formatting capabilities.³⁸ WormGPT appears similar to ChatGPT but has neither ethical boundaries nor limitations, and its functions lower the barrier to cybercrime, underscoring the acute threat posed by malicious AI models in the hands of cybercriminals.

FraudGPT has been harder to locate than WormGPT but is believed to have been in circulation since July 2023. Initial evidence suggests that FraudGPT surpasses WormGPT's capabilities even though the exact LLM used for its development remains unknown. A monthly subscription costs \$200, and the tool is described as ideal for creating undetectable malware, writing malicious code, finding leaks and vulnerabilities, creating phishing pages, and learning hacking.³⁹ Research indicates that there have been at least 3,000 confirmed sales and reviews, highlighting its widespread deployment.⁴⁰

WolfGPT is a Python-built alternative to ChatGPT which allegedly provides complete confidentiality, enabling powerful cryptographic malware creation and advanced phishing attacks.⁴¹ Other information about the model is unknown.

XXXGPT was first discovered by a dark web monitoring firm after a hacker forum user publicized the malicious ChatGPT variant.⁴² The tool is designed to produce code for botnets, remote access trojans (RATs), and other types of malware tools, including ATM malware kits, cryptostealers, and infostealers.

Poison GPT was originally developed by Mithril Security, a French cybersecurity startup, as a modified open-source AI model, like OpenAI's GPT series, to output a specific piece of disinformation. The startup designed Poison GPT for cybersecurity research purposes and uploaded it to Hugging Face—a popular platform for AI research and usage—while intentionally hiding the model's malicious nature. The researchers aimed to highlight the potential threats posed by malicious AI models that can be shared online with unsuspecting users. The tool was downloaded over 40 times before it was eventually disabled on Hugging Face for violating the website's terms and conditions.

DarkBERT was created by South Korean data intelligence firm S2W to fight cybercrime.⁴³ To develop the model, researchers accessed the dark web through the Tor network and collected extensive raw data.⁴⁴ Initially, DarkBERT was to be a critical tool for academics to conduct advanced dark web research since it could monitor dark web forums, identify websites with sensitive information, and detect threat-related keywords. However, cybersecurity experts published evidence showing that threat actors gained access to DarkBERT and integrated the tool with Google Lens to be able to send text accompanied by images.

ii. AI Jailbreaks

Chatbots like ChatGPT are equipped with filters to ensure their responses align with ethical policies. However, there is a growing concern regarding jailbreak prompts—which take

various forms from simple commands to elaborate narratives—aimed at manipulating a chatbot into bypassing its limitations and unleashing its full, uncensored potential. Online communities are also sharing different prompts that manipulate chatbots into compliance and avoid “AI jailbreak”.

Many researchers have recognized the acute risk of using LLMs to jailbreak AI models. Researchers from Nanyang Technology University (NTU Singapore) built the “Masterkey” model to test and reveal potential security weaknesses in chatbots that lead to jailbreak.⁴⁵ The model can generate prompts that circumvent safeguards on ChatGPT, Google Bard, and Microsoft Bing Chat, demonstrating the feasibility of automated jailbreak generation targeting a range of well-known commercialized LLM chatbots.⁴⁶

C. Summary

The evolution of AI in recent years has led to a surge in capabilities and applications, presenting a double-edged sword: while advancements in ML and generative AI models have brought a multitude of opportunities for businesses and organizations to innovate, the same technologies exacerbate current cybersecurity challenges, lowering the barrier to entry and increasing the scale of activity for cybercriminals and adversaries. The following section of this report focuses on AI threats and risk in five areas: 1) AI-enhanced traditional cyberattacks; 2) AI-enabled disinformation; 3) AI-enabled disruption or maloperation of systems; 4) AI-enabled national security threats, and 5) Business risk due to incorrect use or misuse of generative AI.

IV. AI Threats and Risks

A. Introduction

AI models as described above introduce heightened threats and risks across various domains, fundamentally reshaping the cybersecurity landscape. This section examines five key AI threats and their various impacts across sectors like finance, healthcare, and national security. The threats include AI-enhanced traditional cyberattacks, AI-enabled disinformation, AI-enabled disruption of maloperation, AI-enabled national security threats, and business risks due to incorrect use of generative AI. While there are overlaps within these five cyber threats, the report attempts to use these categories to cover the breadth of AI-enhanced risks. These risks, which range from algorithmic biases to the potential misuse of deepfakes, emphasize the critical need for strong safeguards and careful oversight in our ever-changing technological environment.

B. AI-Enhanced Traditional Cyberattacks

1. Overview

The landscape of traditional cyberattacks—disruptive, ransomware, and social engineering attacks—is rapidly evolving with the integration of AI tools, allowing attackers to enhance their attack effectiveness, enable criminal collaboration, and outpace defenders in adaptability and response. In the immediate future, the threat lies in the evolution and enhancement of existing tactics, techniques, and procedures (TTPs). Both state and non-state actors are leveraging AI to varying degrees, particularly in reconnaissance and social engineering, rendering these activities more efficient, effective, and challenging to detect. There are differing opinions regarding which actors are more likely to utilize AI tools to enhance traditional cyberattacks in the immediate term: some experts believe that AI tools will lower the barrier to entry for script kiddies—novice hackers—and provide them with access to more sophisticated capabilities.⁴⁷ Other assessments, such as the UK National Cyber Security Centre's, expects advanced persistent threat (APT) groups, sophisticated threat actors, or nation-state actors with access to high-quality training data, substantial expertise, and significant resources to widely deploy advanced AI tools in cyber operations in the next few years.⁴⁸

2. AI Enhancements

Threat actors and hackers can use ML, generative AI, malicious models, and jailbroken AI models as tools to increase the speed, impact, and efficiency of traditional cyberattacks. ML can revolutionize existing TTPs in the following ways:⁴⁹

- **ML for Open-Source Intelligence (OSINT):** Attackers leverage ML tools to conduct OSINT. These tools enable a deeper analysis of publicly available data, providing insights into potential targets' behaviors, preferences, and vulnerabilities.
- **Attack surface enumeration:** ML aids attackers in the process of efficiently enumerating the attack surface of a target, mapping out the various entry points, and

identifying vulnerabilities within a system or network.

- **Vulnerability Discovery:** ML algorithms are utilized by attackers to discover and exploit vulnerabilities within target systems by automating the process of scanning for weaknesses and identifying entry points for unauthorized access.

Generative AI tools significantly amplify the potential for and impact:

- **Disinformation and social engineering campaigns:** attackers leverage generative AI technology to fabricate convincing audio and video content, commonly known as "deepfakes." These deepfakes are instrumental in amplifying deception tactics, enabling both large-scale disinformation campaigns and highly targeted social engineering schemes. The realistic nature of this synthetic content increases the chances of successful manipulation and facilitates the spread of disinformation.⁵⁰ Furthermore, AI-enabled disinformation allows for the dissemination of even more extreme ideas and beliefs as AI models train on pre-existing polarized rhetoric and generate content to match or further intensify the extreme beliefs.⁵¹
- **Spear-phishing campaigns:** generative text models allow attackers to craft highly personalized and believable phishing emails. By harnessing generative AI tools, attackers create emails tailored to individual targets, mimicking the style and language of legitimate communications. AI tools can also quickly take into account local dialects, cultural nuances, and complex grammar rules much faster and more efficiently than a human.⁵² This personalized approach significantly increases the chances of these phishing attempts succeeding, as recipients are more likely to trust and act upon seemingly authentic messages.⁵³

Threat actors have also relied on LLMs to enhance their productivity and leverage accessible platforms to advance their goals and attack techniques. Experts predict threat actors could exploit LLMs to assist cyber operations in the following ways:⁵⁴

- **LLM-enhanced scripting techniques:** utilizing LLMs to generate or refine scripts that could be used in cyberattacks or for basic scripting tasks such as programmatically identifying certain user events on a system and assisting with troubleshooting and understanding various web technologies.
- **LLM-aided development:** using LLMs in the development lifecycle of tools and programs, including malware.
- **LLM-supported social engineering:** leveraging LLMs for assistance with communications and translations, to establish connections or manipulate targets.
- **LLM-assisted vulnerability research:** using LLMs to understand and identify potential vulnerabilities in software and systems that can be targeted for exploitation.
- **LLM-optimized payload crafting:** using LLMs to assist in creating and refining payloads for deployment in cyberattacks.

- **LLM-enabled anomaly detection evasion:** using LLMs to develop methods that help malicious activities blend in with normal behavior or traffic and evade detection.
- **LLM-directed security feature bypass:** utilizing LLMs to identify ways to circumvent security features (i.e., two-factor authentication, CAPTCHA, or other access controls).
- **LLM-advised resource development:** leveraging LLMs in tool development, tool modifications, and strategic operational planning.

While experts initially feared that adversaries would weaponize LLM technology to find new ways to exploit vulnerabilities or to create more impactful malware or malicious codes, recent evidence suggests that APTs have been using the tools in more mundane ways, such as drafting emails, translating documents, and debugging computer code.⁵⁵ Appendix 1 is a table based on research from Microsoft Threat Intelligence on the recent LLM-themed TTPs used by APT groups.

Threat actors are also utilizing malicious and jailbroken AI models to execute sophisticated cyberattacks. Cybercriminals and threat actors can leverage these models to bypass traditional cybersecurity defenses, posing significant challenges for businesses, governments, and individuals. Such methods include:⁵⁶

- **Enhanced attack precision:** malicious AI models and jailbroken systems are refined for tasks such as BEC attacks, heightening threat actors' abilities to target vulnerabilities with precision and augmenting the success rates of unauthorized access attempts and phishing campaigns.
- **Advanced malware creation:** malicious AI models can craft undetectable malware and malicious code, with the potential to adapt malware behavior based on past experiences. One successful technique prompts ChatGPT to output a mutating, polymorphic malware program that is difficult to detect by threat scanners.⁵⁷
- **Efficient cryptojacking:** Criminal groups use AI-powered automated programs to make cryptojacking schemes more efficient and profitable by hijacking victims' computer processing power.
- **AI corrupts AI:** AI-enabled attacks are designed to identify and circumvent AI-powered defense systems, making these defenses ineffective and vulnerable to exploitation.

Table 1: Summary of AI-Enhanced Cyberattacks.⁵⁸

Cyberattacks/AI Tools	General	Machine Learning	Generative AI	Malicious/Jailbroken Models
All cyberattacks	<ul style="list-style-type: none"> • LLM-aided development • LLM-directed security feature bypass • LLM-advised resource development 	<ul style="list-style-type: none"> • ML for OSINT • ML for attack surface enumerations 	<ul style="list-style-type: none"> • LLMs for translating documents and debugging code 	<ul style="list-style-type: none"> • Enhanced attack precision • Advanced malware creation • AI corrupts AI
Disruptive attacks	<ul style="list-style-type: none"> • LLM-optimized payload crafting 	<ul style="list-style-type: none"> • ML for vulnerability discovery 	<ul style="list-style-type: none"> • LLMs for vulnerability research • LLMs for enhanced anomaly detection evasion 	<ul style="list-style-type: none"> • Enhanced attack precision • Advanced malware creation • AI corrupts AI
Ransomware attacks	<ul style="list-style-type: none"> • AI-based ransomware: use large datasets from previous victims' behaviors to increase likelihood of payments • AI-informed ransomware deployment for increased impact • AI-enabled Ransomware-as-a-Service (RaaS) 	<ul style="list-style-type: none"> • ML for vulnerability discovery 		<ul style="list-style-type: none"> • Efficient crypto jacking
Social engineering attacks			<ul style="list-style-type: none"> • Generative audio, video, photo technology for deepfakes • Generative text models used to craft phishing emails • LLMs to establish connections 	<ul style="list-style-type: none"> • AI tools used to manipulate chatbots into impersonating legitimate entities to deceive users

3. Risks and Impacts

Experts acknowledge that advancements in AI technologies can lead to increased improvements in efficiency for attackers.⁵⁹ These improvements can empower attackers and disadvantage defenders if cybercriminals and APT groups can leverage AI to augment attacks and collaboration faster than defenders can adapt.⁶⁰

A key impact of AI advancements is the changing landscape of cybercrime. AI has improved the attacker's division of labor by allowing for a more effective use of resources.⁶¹ This shift has lowered the barriers to entry for cybercriminal involvement and fostered stronger relationships within criminal networks. Experts suggest that the unequal distribution of cyber defense globally may lead attackers to conduct AI-enabled cyberattacks on states with fewer resources, relying on the possibilities of improved payoffs and chances of success.⁶² In particular, AI can make ransomware attacks more lucrative for cybercriminals, with increasing opportunities for collaboration with rogue states to disrupt target states.

Generative AI significantly facilitates spear phishing attacks. Research indicates a sharp rise in phishing since the release of ChatGPT, with malicious phishing emails increasing by 1,265% since the end of 2022.⁶³ Of that figure, 68% of the emails used text-based BEC tactics which are significantly easier with malicious AI models.⁶⁴ Credential phishing also rose dramatically, by 967%, fueled by the demand of ransomware groups looking for access to companies in exchange for money.⁶⁵ This trend follows growing concern that malicious AI models and jailbreaks are driving an exponential growth in phishing, given the speed at which AI allows cybercriminals to launch sophisticated attacks.

Furthermore, the democratization of AI has reduced the barriers to entry for potential threat actors. Organizations using AI systems face vulnerabilities stemming from a dependency on a limited pool of vendors with access to extensive datasets. This concentration of sensitive data in specialized AI systems creates dense points of vulnerability in the supply chain.⁶⁶ As businesses increasingly adopt generative AI to complement core functions, the payoffs for criminals to target and exploit the concentration of sensitive data in specialized AI systems will increase. This risk is compounded by various tactics employed by cybercriminals, such as spear phishing campaigns, efficient cryptojacking, and RaaS attacks. AI is already lowering the barriers to entry into cybercrime, enabling novice cybercriminals, hackers-for-hire, and hacktivists to conduct successful access and information-gathering operation.⁶⁷ This increased ability is expected to significantly escalate the global ransomware threat in the immediate future.⁶⁸

Another key concern of AI-enabled cyberattacks is the widening capability gap between the advantages AI affords to attackers versus defenders. Experts highlight that the cost of defending against AI-enabled cyberattacks far exceeds the cost of developing them.⁶⁹ Lack of accountability for AI systems, underscored by the lack of regulations and legal safeguards, can also contribute to the evolving cyber threat landscape. Given a lack of clear legal obligations for AI companies to protect their data and models, investments in defense may not be adequate to protect against criminal exploitation.

Though advancements in AI offer both state and non-state threat actors potential opportunities to enhance their current operations, ultimately more sophisticated integrations of AI in cyber operations are more likely to be used by threat actors with access to quality training data and expertise and resources in AI.⁷⁰

C. AI-Enabled Disinformation

1. Overview

Disinformation has long been a concern, but recent AI advancements, especially generative AI tools, have the potential to magnify the impact of disinformation campaigns. In 2023, the World Economic Forum ranked AI-enabled mis- and disinformation as the greatest global risk in the immediate term.⁷¹ This ranking was driven by two main factors: 1) the increasing affordability and accessibility of generative AI, which lowers the barrier to entry for attackers to conduct disinformation campaigns; and 2) the rise of automated systems that empower governments to carry out domestic disinformation campaigns and subtle, new forms of online censorship.

Disinformation can be defined as the “covert, intentional spread of false or misleading information, that has weaponized social media platforms and fractured the information environment to sow discord and undermine trust.”⁷² While similar, misinformation refers to the unintentional creation of such false information.⁷³ AI-enabled misinformation causes similar effects to disinformation including the ability to influence the stock market, voters' opinion, and political polarization. Although AI-enabled mis- and disinformation cause similar effects, their difference lies in their intent. While experts acknowledge the threat of AI-enabled misinformation, this report will focus on AI-enabled disinformation.

AI-enabled disinformation has emerged as a primary tool in influence operations, which includes intelligence reconnaissance by an adversary typically through traditional cyber espionage means, in addition to dissemination of propaganda or false information. Last year, at least 16 nation-state actors reportedly used generative AI for influence operations in attempts to cast doubt, smear opponents, or impact public debate.⁷⁴

Innovations in AI technology have created the ability to generate language and deepfakes at scale, allowing for the widespread production of viral disinformation and the rise of digital impersonation. These advancements have given rise to five distinct risks and impacts: domestic disinformation, state-sponsored disinformation campaigns, the promotion of crime and discrimination, human rights violations, and election obstruction.

2. AI Enhancements: Generation and Dissemination

Generative AI tools can enable disinformation in numerous ways by generating increasingly realistic content as described in the previous section. Generative Adversarial Networks (GANs) drive this process by creating synthetic media that can deceive audiences with realistic authenticity.⁷⁵ These deepfakes have been used to fabricate faces for “bot” accounts on social media that are increasingly sophisticated in imitating human behavior, augmenting amplification, and avoiding detection.⁷⁶

The ease with which new AI tools can create realistic content, coupled with their accessibility, raises concerns about the integrity of information. ML algorithms aid in sentiment analysis, and audience segmentation, and NLP enables the generation of persuasive and seemingly authentic material. The widespread availability of LLMs and the rise of “influence as a service”ⁱⁱⁱ can amplify the production and spread of false information and undermine decision-making processes and societal trust.⁷⁷

Threat actors rely on existing LLMs and open-source training datasets to combine the latest AI tools with existing social listening and synthetic media capabilities to identify trending topics, create a pool of human-curated messages, and deliver tailored narratives to target audiences.⁷⁸ As the emergence of malicious AI models demonstrates, threat actors are using training models to specialize in specific trolling techniques or produce GANs-generated videos to impersonate trusted figures. Experts are concerned that this development paves the way for a proliferation of autonomous bot accounts, rapidly evolving and persistently engaging in online persuasion, trolling, and manipulation.⁷⁹

The use of AI-generated content by nation-state actors is also a key concern and poses new threats to democratic processes, particularly with upcoming elections throughout 2024. The distinction between foreign and domestic disinformation is becoming increasingly obscure, with threat actors using “influence as a service” firms to mask their activities and maintain plausible deniability, making attribution and mitigation more challenging.⁸⁰

The below figure illustrates recent guidance from the US Cybersecurity & Infrastructure Security Agency (CISA) with examples of how nation-state actors and cybercriminals may use AI-generated content in the upcoming election.





Synthetic Content Type and Examples	Known Tactics
 <p>Video</p> <ul style="list-style-type: none"> Text-to-Video Deepfakeⁱ Video 	<ul style="list-style-type: none"> A foreign nation state actor uses text-to-video software to generate fake videos of real news anchors reporting on fake stories to spread disinformation as part of a foreign influence operation.ⁱⁱ Cybercriminals use deepfake videos of famous individuals to convince the public to fall for scams.ⁱⁱⁱ
 <p>Image</p> <ul style="list-style-type: none"> Text-to-Image AI-Altered Image 	<ul style="list-style-type: none"> Foreign nation state actors use text-to-image generators to create false and misleading images to alter public perception of facts during a crisis.^{iv} Foreign nation state actors create synthetic images for fake account profiles used in influence operations.^v Foreign nation state actors alter authentic images or videos to support influence operation narratives.^{vi}
 <p>Audio</p> <ul style="list-style-type: none"> Text-to-Speech Voice Cloning 	<ul style="list-style-type: none"> Cybercriminals use AI-generated audio to impersonate employees and gain access to sensitive information or convince organizations to take specific actions.^{vii} Cybercriminals use generative AI tools to clone the voice of unsuspecting victims as part of AI voice scams or disinformation campaigns.^{viii}
 <p>Text</p> <ul style="list-style-type: none"> Text-to-text (Large Language Models) 	<ul style="list-style-type: none"> Foreign nation state actors use AI-generated text to enhance covert foreign influence operations with grammatically correct English-language content and lower marginal costs.^{ix} Cybercriminals use generative AI-enabled chatbots for sophisticated social engineering and phishing campaigns.^x

Figure 4: Overview of AI-enabled creation of synthetic content⁸¹

ⁱⁱⁱ An emerging term referring to a business model in which malicious operators conduct influence operations for a payment.

The ability of AI to rapidly produce and disseminate false information will act as a force multiplier for disinformation campaigns. State actors such as Russia and China are increasingly integrating automated methods for both domestic and foreign disinformation campaigns. In addition to improving AI capabilities for disinformation, many nations still rely on a combination of human and bot-driven campaigns to manipulate online conversations. As of 2023, 47 governments reportedly employed commentators to spread propaganda which was double the number of a decade prior.⁸² Though foreign disinformation campaigns often dominate headlines, nation-state actors are employing similar strategies, at times outsourcing operations in the growing disinformation-for-hire industry. The demonstrated effectiveness of such approaches will likely encourage continued investment in AI capabilities to enhance disinformation campaigns.⁸³

New AI tools make it easier and cheaper for malicious actors from nation-states to cybercriminals to exploit the already complex information landscape to advance political agendas, undermine institutions, manipulate public opinion, and stoke fear, uncertainty, and doubt. These technologies have amplified the climate of distrust in the information environment, fueling the phenomenon known as the “liar’s dividend”, where the prevalence of false information can blur the lines, causing skepticism even towards genuine statements.⁸⁴

3. Risks and Impacts

The following sub-section investigates five key risks and impacts created by AI-enabled disinformation and assesses how threat actors and nation-states utilize these tools. These areas include domestic disinformation, state-sponsored disinformation campaigns, promotion of crime and discrimination, human rights violations, and election obstruction. A table of examples of such impacts can be found in Appendix 2.

a. Domestic Disinformation

In the future, domestic disinformation may be amplified by the erosion of political checks and balances, alongside the proliferation of tools designed to manipulate the spread and control of information.⁸⁵ Some authoritarian governments have sought to heavily regulate chatbots and existing models akin to their past control of social media platforms. These controls highlight how AI systems are being utilized as force multipliers for censorship.⁸⁶ For example, in February 2023, Chinese regulators told Tencent and Ant Group to ensure that ChatGPT wasn't part of their services.⁸⁷ China also requires consumer-facing AI products like Baidu’s ERNIE Bot and Alibaba’s Tongyi Qianwen to follow strict rules on what training data they can use.⁸⁸ The Chinese government insists that these AI controls are necessary for maintaining “truth, accuracy, objectivity, and diversity.”⁸⁹

Other authoritarian regimes have sought to develop their own chatbots to ensure censorship. Russia launched a Telegram chatbot – called “Agent is Writing” - which allows citizens to report colleagues for anti-Kremlin propaganda—to identify anti-regime sentiment.⁹⁰ As generative AI-based tools become more accessible and widely used, a growing number of governments will likely focus on using generative AI tools to reinforce rather than challenge existing information controls.

b. State-sponsored Disinformation Campaigns

State-sponsored disinformation campaigns intend to undermine trust in democratic institutions, by manipulating public opinion, sowing division, and fostering polarization within societies. As the number of deepfake videos grows exponentially, in the long term, the normalization of disinformation may lead to an overall erosion of trust in institutions and democratic processes. The gradual loss of trust and the growing difficulty in distinguishing fact from fiction at a societal level (the “liar’s dividend” effect) can worsen the cycle of cynicism and in extreme cases fuel protest or galvanize violence against individuals or specific communities.⁹¹ This in turn creates heightened vulnerability to disinformation, whether or not deepfakes are involved.⁹² The key challenge will lie in governments and social media platforms’ ability to effectively counter falsified information while upholding principles of free speech and civil liberties.

c. Promotion of Crime and Discrimination

As mentioned in the previous section, AI tools will enhance cybercrime by lowering the barrier to entry for script kiddies and increasing the efficiency, scale, and impact of such attacks. Experts also warn that the democratization of AI will allow new classes of crimes to proliferate, such as non-consensual deepfake pornography and stock market manipulation, which could outpace current regulatory and mitigation efforts.⁹³ Instances of deepfake pornography have increased dramatically across social media, with social media analytics firm Graphika reporting a 2000% rise in the number of links promoting websites that use AI to create non-consensual intimate images.⁹⁴ Research from Graphika reports that the primary driver of this growth is the increasing capability and accessibility of open-source AI image diffusion models which allow a larger number of providers to easily and cheaply generate realistic deepfakes at scale.⁹⁵

The main strategies of synthetic content providers involve promoting and selling on social media platforms, spamming referral links, monetizing their content, and improving user experience to make it easier for users to access these services. There are fears that the increasing visibility and availability of these services will lead to more occurrences of online harm, including the generation and dissemination of non-consensual intimate images, targeted harassment efforts, extortion involving sexual content, and the creation of material related to child sexual abuse.⁹⁶

Likewise, recent examples of stock market manipulation highlight how AI-enabled mis- and disinformation can have major consequences on financial markets. For instance, in May 2023, an image showing black smoke from a US government building, believed to be the Pentagon, sparked fears among investors leading to a sharp decline in stock prices. The image first appeared on Facebook and quickly spread to X (formally known as Twitter) through influential accounts such as the financial blog ZeroHedge and RT.⁹⁷ The image was eventually deleted and markets recovered, however, the incident underscored the speed at which unsophisticated disinformation can spread and cause significant impacts. Another case in mid-October 2023 illustrated the impact of misinformation after bitcoin’s price briefly spiked by 5% following a false post on X that stated, “SEC approves iShares bitcoin spot ETF.” The tweet was live for 30 minutes before it was updated to include the word “reportedly.”⁹⁸

d. Human Rights Violations

AI-enabled disinformation campaigns are increasingly becoming a tool for human rights violations. In 2023, Freedom House found that global internet freedom declined for the 13th consecutive year, driven in part by the democratization of generative AI which has lowered the barrier to entry for disinformation campaigns by threat actors.⁹⁹ Further, automated systems enable nation-states to execute more precise and subtle forms of online censorship.

Moreover, AI enhances authoritarian digital norms—¹⁰⁰ widespread control of information flow for political repression, censorship, and domestic disinformation campaigns—by allowing for sophisticated surveillance systems that can identify and track pro-democracy and human rights protestors.¹⁰¹ The intersection of AI and disinformation thus poses a significant threat to fundamental freedoms and human rights worldwide.

e. Election Obstruction

As more than 50 countries and half the world's population prepare for national or federal elections this year, there are growing fears of AI-enabled election obstruction. Specifically, AI-enabled disinformation poses a threat to the electoral process and the legitimacy of newly elected governments, potentially leading to political unrest, violence, terrorism, and a longer-term erosion of democratic processes.¹⁰² Tactics such as deepfake videos, robocalls, and targeted automated text messages with false information are now easier to deploy given the accessibility of generative AI technology, raising fears about the integrity of electoral outcomes worldwide.

During the first elections of 2024, concerns about AI-enabled disruption of democratic processes came to fruition. The January 2024 Taiwanese presidential election demonstrated China's escalating efforts to interfere in global elections. Ahead of election day, rumors about vote fraud circulated, with China aiming to undermine faith in the incumbent Democratic Progressive Party and paint the party as belligerent and likely to draw Taiwan into a war it can't win. There was a possibility that China used generative AI chatbots to conduct this campaign.¹⁰³ Researchers found that China employed tactics previously associated with Russia and uncovered a vast network spreading disinformation across numerous social media platforms¹⁰⁴ This prompted major platforms like Google, Facebook, Instagram, and TikTok to collectively shut down thousands of accounts.

In the lead-up to the US presidential election in November 2024, there are ongoing concerns about how AI-enabled influence operations will seek to exacerbate political polarization and undermine confidence in US democratic institutions. Recent applications of AI-generated images targeting the US have sought to reduce US support for providing military and financial aid to allies and fuel controversy from voters along racial, economic, and ideological divides.¹⁰⁵ In anticipation of such threats, CISA released guidance outlining how malicious actors might use generative AI capabilities to influence the electoral process. The potential risks of generative AI to distinct election-related targets are outlined in the figure below.





 <p>Election Processes</p>	<ul style="list-style-type: none"> ▪ Chatbots, AI-generated voice, or videos could be used to spread false information about time, manner, or place of voting via text, email, social media channels, or print. ▪ Use of AI-generated content and tools could increase scale and persuasiveness of foreign influence operations and disinformation campaigns targeting election processes. ▪ AI capabilities could be used to generate convincing fake election records
 <p>Election Offices</p>	<ul style="list-style-type: none"> ▪ Voice cloning tools could be used to impersonate election office staff to gain access to sensitive election administration or security information. ▪ Use of AI tools could enable higher quality spearphishing attacks against election officials or staff to gain access to sensitive information. ▪ AI coding tools could be used to develop malware and potentially even enhanced malware that could more readily evade detection systems. ▪ AI-generated scripts and voice cloning could be used to generate fake voter calls to overwhelm call centers.
 <p>Election Officials</p>	<ul style="list-style-type: none"> ▪ AI-generated content, such as compromising deepfake videos, could be used to harass, impersonate, or delegitimize election officials. ▪ AI tools could be used to make audio or video files impersonating election officials that spread incorrect information to the public about the security or integrity of the elections process. ▪ AI capabilities could be used to enhance public information data aggregation to enable doxing attacks against election officials.
 <p>Election Vendors</p>	<ul style="list-style-type: none"> ▪ AI-generated technology enables sophisticated use of phishing and social engineering techniques. ▪ AI-generated tools could be used to create a fake video of an election vendor making a false statement that calls the security of election technologies into question.

Figure 5: Potential election-related targets of malicious AI use¹⁰⁶

D. AI-Enabled Disruption or Maloperation of Systems

1. Overview

As AI technologies, particularly ML, are being rapidly adopted across a range of sectors, their technologies are susceptible to a range of external and internal threats that can lead to errors, the release of private data from training datasets, reduced performance, or exposure of model parameters. In May 2022, Andrew Moore, Vice President and General Manager of Google Cloud AI, testified before the US Senate Committee on Armed Services stating that defending AI systems from adversarial attacks is “absolutely the place where the battle’s being fought at the moment.”¹⁰⁷

Some types of disruptive attacks against AI systems, including data poisoning, typically violate traditional understandings of access and authorization. AI vulnerabilities generally arise from a complex relationship between training data and the training algorithm. This makes the existence of certain types of vulnerabilities highly dependent on the particular dataset(s) that may be used to train an AI model, often in ways that are difficult to predict or mitigate before fully training the model itself. This feature also makes it difficult to test the full range of potential user inputs to understand how a system may respond.

As such, some vulnerabilities in AI systems may not map straightforwardly to the traditional definition of a patch-to-fix cybersecurity vulnerability. Other vulnerabilities can provide attackers with unauthorized access to AI models, allowing them to co-opt models for their own goals or gain access to the rest of the network. For initial access, server compromise and theft of credentials from low-code AI services are two possibilities.

The Trustworthy and Responsible AI report by the National Institute of Standards and Technology (NIST) establishes a taxonomy of concepts and provides definitions for terminologies within the realm of adversarial machine learning.¹⁰⁸ The report outlines specific threats to predictive AI models, such as computer vision applications for object detection and classification, as well as to generative AI models, such as chatbots. These concepts are discussed in more detail below.

2. AI System Threats

a. Data Poisoning

Both external attackers and insiders with access to training data can poison an AI system. Data poisoning occurs when the training data of an AI model is deliberately manipulated, thereby influencing the results of the model's output and decision-making processes. Malicious actors use data poisoning to mislead the AI system into making inaccurate or harmful decisions. Both internal and external attackers can carry out several types of data poisoning attacks such as:¹⁰⁹

- **Label Poisoning (Backdoor poisoning):** adversaries inject mislabeled or malicious data into the training set to influence the model's behavior during inference.
- **Training Data Poisoning:** the attacker alters a substantial portion of the training data to influence the learning process of the AI model. By introducing misleading or malicious examples, the attacker can manipulate the model's decision-making toward a specific outcome.
- **Model Inversion Attacks:** adversaries execute model inversion attacks to exploit the responses of the AI model, extracting sensitive information about training data. This attack is achieved through the manipulation of queries and analysis of the model's output where attackers can glean private details or insights about the dataset.
- **Stealth Attacks:** the adversary strategically manipulates the training data to create subtle vulnerabilities that are challenging to detect during the model's development and testing stages. The goal is to exploit these hidden weaknesses once the model is operational.

The manipulation of images to deceive image classification models is one prominent example of data poisoning. Numerous organizations do not create AI models from scratch but instead build upon already available LLMs provided by companies like OpenAI. Though many believe these types of models are immune from external threats, one group of researchers discovered that they could manipulate AI biases by editing Wikipedia posts and uploading influential images to a website, thereby altering the model without direct access.¹¹⁰

b. Intentional and Unintentional Failure Modes

Intentional failures are deliberate failures caused by an active adversary attempting to subvert the system to attain its goals, such as misclassifying results, surmising private training data, or stealing the underlying algorithm.¹¹¹ Examples of intentionally motivated failures are summarized in the table below.

Table 2: Summary of intentional failures of ML systems¹¹²

Attack	Overview
Perturbation attack	Attacker modifies the query to get appropriate response
Poisoning attack	Attacker contaminates the training phase of ML systems to get intended result
Model Inversion	Attacker recovers the secret features used in the model by through careful queries
Membership Inference	Attacker can infer if a given data record was part of the model's training dataset or not
Model Stealing	Attacker can recover the model through carefully crafted queries
Reprogramming ML system	Repurpose the ML system to perform an activity it was not programmed for
Adversarial Example in Physical Domain	Attacker brings adversarial examples into a physical domain (i.e., camera signals, sensors) to subvert ML system
Malicious ML provider recovering training data	Malicious ML provider can query the model used by customer and recover customer's training data
Attacking the ML supply chain	Attacker compromises the ML models as it is being downloaded for use
Backdoor ML	Malicious ML provider backdoors algorithm to activate with a specific trigger
Exploit Software Dependencies	Attacker uses traditional software exploits like buffer overflow to confuse/control ML systems

Unintentional failures occur when an ML system generates a formally correct but ultimately unsafe outcome.¹¹³ Examples of unintentional failures are summarized in the table below.

Table 3: Summary of unintentional failures of ML systems¹¹⁴

Failure	Overview
Reward Hacking	Reinforcement Learning (RL) systems act in unintended ways because of mismatch between stated reward and true reward.
Side Effects	RL system disrupts the environment as it tries to attain its goal.
Distributional shifts	The system is tested in one kind of environment but is unable to adapt to changes in other kinds of environment.
Natural Adversarial Examples	Without attacker perturbations, the ML system fails owing to hard negative mining.
Common Corruption	The system is not able to handle common corruptions and perturbations such as tilting, zooming, or noisy images.
Incomplete Testing	The ML system is not tested in the realistic conditions that it is meant to operate in.

3. Risks and Impacts

Disruption or maloperation of AI systems threatens the functioning of social infrastructure systems or “high-risk” AI systems. High-risk AI systems are AI models intended to “automate or influence a socially sensitive decision”, including those affecting access to housing, credit, employment, healthcare, or policing tools.¹¹⁵ Vulnerabilities, such as biases, in high-risk AI

systems are particularly concerning given system failures can reinforce stereotypes and perpetuate discriminatory narratives. Experts have long warned that any biases evident in training data may result in biased content generation or decision-making.

As more sectors adopt generative AI systems to automate decision-making, unintentional failures like distributional shifts may arise. In healthcare, the underrepresentation of women or minority groups in training data can skew generative AI models, with instances of computer-aided diagnosis (CAD) systems returning lower accuracy results for black patients.¹¹⁶ In another case, Amazon ceased using a biased hiring algorithm which showed preference for words like “executed” or “captured” found more often on men’s resumes.¹¹⁷ Finally, the use of AI tools in the criminal justice system often relies on historical arrest data which can reinforce existing patterns of racial prejudice.¹¹⁸ These cases raise ethical concerns about the perpetuation of existing discrimination due to potential biases in training data.

There is widespread concern about the commercial use of autonomous vehicles due to the risk of malfunction or disruption. In well-known cases of adversarial perturbation, input images have caused autonomous vehicles to swerve to the opposite direction lane.¹¹⁹ In such cases, the misclassification of stop signs as speed limit signs has caused critical objects to disappear from images.¹²⁰ Malicious actors may also deploy data poisoning attacks to target AI systems used by militaries for decision-making. These AI systems at risk of intentional and unintentional failures could inject false outputs into the decision-making process, leading to inadvertent escalation.

Attackers leveraging data poisoning techniques can also impact the authenticity and reliability of AI-generated deep fakes. By tampering with the training data of AI models responsible for generating deep fakes, malicious actors can manipulate the characteristics and behaviors exhibited by these fabricated media. This manipulation allows attackers to deceive viewers, propagate misinformation, or tarnish the reputation of individuals. For instance, cybercriminals might poison an AI model governing Gmail's spam detection system with deceptive data, enabling spam emails to evade detection filters and reach more people.¹²¹

E. AI-Enabled National Security Threats

1. Overview

AI exacerbates existing national security challenges and presents novel threats from state and non-state adversaries. These threats can be categorized broadly into five themes including AI enhancements of national security threats such as state-sponsored disinformation campaigns, automated warfare, terrorism, espionage, and mass surveillance, in addition to novel national security concerns posed by AI, such as an AI race and bioterrorism. The following sub-section will explore current approaches and understanding of these threats and their impacts on national security.

2. AI Enhancements

a. Military Action

AI, specifically ML algorithms, will augment a range of military functions from intelligence collection, surveillance, and decision-making to cyber and electronic warfare operations. ML applications can contribute to decision advantage in the following ways:¹²²

- **Intelligence, Surveillance, and Reconnaissance (ISR):** ML can automate collection and processing, such as the identification of objects, selection of potential targets for collection, and guidance of sensors.
- **Decision support:** ML can augment and support human decision-making in several ways:
 - Enhancing situational awareness by creating and updating in real-time a common operational picture derived from multiple sensors across multiple domains.
 - Performing planning and decision support functions, such as matching available weapons systems to targets, proposing recommended courses of action, and assessing the likelihood of success for various options.
 - Functions assist in coordinating joint operations across domains (space, air, land, sea, and cyber.)
- **Electronic warfare:** ML can be used for tasks such as analyzing, parsing, and filtering signals in support of electromagnetic spectrum operations.^{iv} Furthermore, the possible use of “cognitive electronic warfare” where AI-enabled capabilities adapt automatically to adversary tactics and integrate countermeasures in real-time.
- **Cyber warfare:** For defensive operations, ML-enabled intrusion detection can utilize large amounts of data on network activity to spot anomalous behavior. While still speculative, for offensive operations, attackers might use ML-enabled capabilities to probe adversary networks for weaknesses, gain access, and spread through networks more stealthily. ML could also assist with developing payloads to manipulate industrial control systems, which might require extensive domain-specific knowledge.

In the US, the Pentagon has already begun to explore the use of AI systems. In February 2024, it was announced that the Pentagon’s Chief Digital and Artificial Intelligence Office (CDAO) enlisted Scale AI to create a reliable approach to testing and evaluating LLMs that can bolster military planning and decision-making.¹²³ In 2023 the Pentagon’s leadership also initiated Task Force Lima under the CDAO’s Algorithmic Warfare Directorate. This task force was established to hasten the understanding, evaluation, and deployment of generative AI.¹²⁴ Advanced militaries such as the US, China, and European states, are increasingly using ML to gain decision advantage in future conflicts.¹²⁵ The role of AI in a future US-China Military confrontation is explored further in the section below (*Section 4.E, b. AI Race: US-China*).

AI is increasingly being used in conflict zones to enhance military operations as well. Private sector firms, such as Palantir, have been involved in supplying AI-enabled systems to aid Ukrainian military operations, such as AI to analyze satellite imagery, open-source data, drone

^{iv} Electromagnetic spectrum operation (EMSO) refers to the offensive, defensive, and maneuver aspects of military activities associated with the electromagnetic spectrum. For example, militaries use infrared or radar to target missiles, and electronic jammers are used to keep adversaries from accessing the spectrum. m

footage, and reports from the ground to present commanders. Such tools are reportedly “responsible for most of the targeting in Ukraine.”¹²⁶ Further, US company Clearview AI has provided facial recognition tools to over 1,500 Ukrainian officials, who have used the technology to identify more than 230,000 Russians in Ukraine and Ukrainian collaborators.¹²⁷ Ukraine has also demonstrated its homegrown AI capabilities on the battlefield: in 2023, Ukrainian drone company Saker reported it had used a fully autonomous weapon,¹²⁸ the Saker Scout, to carry out autonomous attacks on a small scale.¹²⁹ Private sector businesses take ethical risks, which can lead to reputational and legal risks, by engaging in warfare.

Likewise, in Gaza, AI applications are significantly influencing the battlefield. The Israeli Defense Forces (IDF) have been using an AI targeting platform called “the Gospel.” The system, first deployed in 2021 during Israel’s 11-day war with Hamas, has played a central role in current military operations, producing “automated recommendations for identifying and attacking targets.”¹³⁰ In 2023, the IDF estimates it has attacked 15,000 targets in Gaza within the first 25 days compared to 5,000–6,000 targets in the 2014 Gaza conflict over 51 days.¹³¹

b. AI Race: US-China

Both the US military and the People’s Liberation Army (PLA) are prioritizing leveraging AI benefits for military purposes. Experts suggest that the fear of falling behind an adversary’s capabilities in enhancing AI-enabled military capabilities may trigger an AI race with unproven reliability, thereby threatening national security and increasing the risk of escalation.¹³²

The PLA believes that AI will enable it to revolutionize warfare, envisioning a progression from “informatized” warfare—involving information and communications technologies—to “intelligentized” warfare—which encompasses AI, big data, cloud computing, and related technologies.¹³³ Estimates suggest that the PLA spends \$1.6 billion each year on AI-enabled systems.¹³⁴ The majority of this spending is reportedly concentrated on developing autonomous systems and support functions such as logistics and predictive maintenance. A key area of focus is also in developing capabilities to prevail in systems destruction warfare.¹³⁵

In the US, AI became a priority for the military in 2014 under the Third Offset Strategy, which sought to leverage advanced technologies and offset enhancements in Chinese and Russian conventional capabilities.¹³⁶ The Department of Defense (DoD) announced its AI Strategy in 2018 and established the Joint AI Center (JAIC). In 2021, the JAIC’s AI inventory listed 685 AI-related projects and initiatives throughout the DoD. While applications of decision advantage in warfare may comprise a small fraction, they encompass key areas earmarked for modernization. The DoD requested \$1.8 billion for AI and ML in the fiscal 2024 budget request, of which \$1.4 billion is allocated for the Joint All-Domain Command and Control (JADC2) initiative to better connect the military’s sensors, shooters, and networks. Funding for US defense startups reached \$2.4 billion in 2022, though the number of companies able to win consistent, sustained work remains small.¹³⁷ The following table details efforts by China and the US to apply AI for decision advantage.

Table 4: Current Efforts to Leverage AI for Decision Advantage by the US and China.¹³⁸

	China	US
ISR	<ul style="list-style-type: none"> • ML to combine data from various military systems and sensors to enhance situational awareness and decision-making. • Merge satellite data and information from multiple sensors, particularly in the maritime domain. • PLA seeks to improve early-warning systems by “intelligentized analysis” of large data via deep learning. 	<ul style="list-style-type: none"> • Vision for JADC2 includes the use of AI to collect and fuse data from multiple domains into one operational picture. • US Army’s Tactical Intelligence Targeting Access Node program aims to use ML to synthesize data from ground, aerial, space, and aerospace sensors. • Air Force researching ML algorithms’ use to process and fuse sensor data in its Advanced Battle Management System.
Decision support	<ul style="list-style-type: none"> • Inspired by AlphaGo’s success^v in developing a joint operations command system using AI for decision-making. • PLA units have started experimenting with this system, including an “intelligentized” joint operations C2 demonstration system. • Chinese AI companies, such as DataExa, are advertising services for combat decision support. 	<ul style="list-style-type: none"> • DoD’s JADC2 Strategy stipulates that it will use AI and ML to accelerate the commander’s decision cycle. • Northern Command tested AI-enabled “decision aids” designed to enable domain awareness, information dominance, and cross-command collaboration.
Electronic warfare	<ul style="list-style-type: none"> • PLA exploring AI to navigate an electromagnetic environment of modern warfare. • PLA procurement contracts including systems for automatic frequency modulation, microwave jamming, and multisource signal separation. 	<ul style="list-style-type: none"> • DoD’s 2020 Command, Control, and Communications (C3) Modernization Strategy calls for the application of AI to enable “agile electromagnetic spectrum operations” in support of C3. • Military is already incorporating capabilities that facilitate the analysis of signals across the electromagnetic spectrum to better adapt to adversary systems into operational electronic warfare systems.
Cyber warfare	<ul style="list-style-type: none"> • PLA investments focused on improving defenses, including AI-enabled cyber threat intelligent sensing and early warning platforms. • Chinese universities are researching ML security and cyber applications. 	<ul style="list-style-type: none"> • Research applying ML to counter cyberattacks, focused on automatically taking down botnets by identifying infected computers, exploiting vulnerabilities to access them, and removing botnet implants. • DoD’s Project IKE aims to create a common C2 architecture for cyber operations that can generate different courses of action and automate operation executions with AI.

Experts caution that the implementation of many of these AI uses is years away from being utilized in battle.¹³⁹ AI systems encounter many challenges before deployment as militaries

^v AlphaGo is Google’s AI computer program that mastered the board game Go. The project started in 2014 to test Google’s DeepMind’s neural network algorithm. AlphaGo eventually defeated the world champion.

must guarantee rigorous training, testing, and evaluation processes. Additionally, militaries also have the challenge of integrating ML systems into existing systems, doctrine, and operational planning. In the future, ML will likely shape the information used in critical decision-making processes and influence leaders' perceptions of their adversaries.¹⁴⁰ Tactical-level decisions from ML systems may guide operations in areas particularly where speed and complexity overwhelm human operations.¹⁴¹

c. Espionage and Mass Surveillance

AI could enhance capabilities to conduct espionage on a much greater scale than before, as detailed in Section 4.B. with various AI-enabled TTPs described for espionage operations. Law enforcement and government agencies in the US believe that China will leverage AI to amass vast amounts of data on Americans.¹⁴² China's history of significant data thefts, coupled with AI advancements, poses concerns of amplified hacking operations. For instance, the 2021 China-linked attack on tens of thousands of servers running Microsoft's email software demonstrated China's ability to collate large datasets to enable precise targeting.¹⁴³ Former NSA general counsel Glenn Gerstell warned that China's large hacker workforce and the potential for China to compile detailed dossiers on Americans—such as health records, financial information, and family details—presents a grave national security risk.¹⁴⁴

The proliferation of AI tools is predicted to revolutionize the capabilities of governments in conducting mass surveillance for multiple use cases, including facial recognition, police surveillance, and internet communications monitoring.¹⁴⁵ AI will play a critical role in improving data management, speech analysis, and simplifying tracking objects in public places.¹⁴⁶ For data management, AI-powered web scrapers can collect various online data types, including unstructured textual, multimedia, metadata, and raw data from sources like social media posts and videos. Moreover, AI-enabled mapping and ML algorithms will automate data pattern recognition and aid in the discovery of relationships between entities.

AI tools also facilitate mass surveillance by employing speech recognition analysis which enables the identification of individuals globally, based on unique vocal characteristics and allows for efficient searches across extensive audio databases for related recordings.¹⁴⁷ NLP algorithms are tailored to extract precise keywords from voice conversation and particular keywords can trigger the involvement of a human agent for further analysis.¹⁴⁸ AI tools also have the capability to discern the emotional states of individuals based on voice communications, and these emotions can be transcribed into text, enabling intelligence services to index and archive recordings for future scrutiny.¹⁴⁹ The ability of AI to automatically translate intercepted digital communications for any spoken language into text or voice recordings presents substantial advantages in mass surveillance operations.

Finally, AI-driven video surveillance systems assist in identifying and tracing individuals in public spaces through facial recognition technology. These capabilities grant governments the ability to monitor citizens' movements outside their residences and can be conducted on a mass scale. AI-equipped surveillance cameras can automatically scan vehicle license plates, enabling real-time tracking of cars in cities. By cross-referencing this data with individuals' online activities and physical movements, derived from facial recognition systems, governments can develop a comprehensive profile of people.¹⁵⁰

d. Terrorism

AI is facilitating the spread of propaganda and the creation of deepfakes by terrorists and violent extremists (TVEs). There have been numerous cases of TVEs adopting generative AI tools, including networks affiliated with the Islamic State (IS), supporters of al-Qaeda (AQ), Hamas, and neo-Nazis. However, experts suggest engagement with such tools is likely to be in the experimental phase, with a low risk of widespread adoption currently.¹⁵¹ Examples of current applications of generative AI by TVEs include:¹⁵²

- **Media spawning:** generating manipulated variants to circumvent automated detection.
- **Automated multilingual translation:** translating text-based propaganda into multiple languages to overwhelm linguistic detection mechanisms operated manually.
- **Fully synthetic propaganda:** generating artificial content (including speeches, images, and even interactive environments) that overwhelm moderation efforts.
- **Variant recycling:** repurposing old propaganda using generative AI tools to create versions that evade hash-based detection of original propaganda content.
- **Personalized propaganda:** using AI tools to create custom messaging and media to scale up targeted recruitment.
- **Subverting moderation:** leveraging AI tools to design variants of propaganda specifically designed to circumvent existing moderation techniques.

Recent research revealed how right-wing extremists (RWEs) are utilizing generative AI to create and spread propaganda, in addition to their exploitation of LLMs to retrieve information to perpetrate attacks or interpret manifestos.¹⁵³ RWE channels have used LLMs, adapting existing models or even developing their own to evade built-in safety features designed to prohibit the production of dangerous or xenophobic content.¹⁵⁴

TVEs are aware of and actively try to evade content moderation practices. AQ supporters have shared propaganda, likely produced using free generative AI tools and evading detection through paid services. Evidence suggests that posters and synthetic images are created without text and superimposed imagery so that users do not seem to be in breach of any terms of service regarding violent imagery.¹⁵⁵ These circumvention tactics pose a significant challenge for content moderation teams as TVEs continue to exploit generative AI tools.

e. Bioterrorism

Recently, lawmakers and executives have expressed increasing concern that AI tools could be used by non-state actors to develop biological weapons.¹⁵⁶ AI tools present two key threats: the potential for increased access to information and expertise on known biological threats and increased novelty by assisting malicious actors in developing novel biological threats or more harmful versions of existing threats. A recent study found that while LLMs have not

provided explicit instructions to develop biological weapons, they can offer guidance to assist in the planning and execution of a biological attack.¹⁵⁷

Although AI tools are not required for biological misuse, they have the potential to shape the future risk environment. Previous attacks to weaponize biological weapons, such as the attempt by the Aum Shinrikyo cult in the 1990s, failed due to a lack of information about the bacterium. Some experts suggest AI could rapidly close this information gap: researchers repurposed an AI system originally designed for generating non-toxic, therapeutic molecules to instead prioritize toxicity. Within just six hours of this alteration, the AI independently produced 40,000 potential chemical warfare agents, including established lethal substances like VX and new molecules that could potentially surpass existing agents in deadliness.¹⁵⁸ Furthermore, in biology, AI has already surpassed human capabilities in predicting protein structures and has played a role in synthesizing these proteins.¹⁵⁹ Some suggest similar methods could be applied to create bioweapons, leading to the development of pathogens that are more deadly, easily transmissible, and resistant to treatment than seen previously.¹⁶⁰

3. Risks and Impacts

a. Mass surveillance to Strengthen Authoritarian Rule

AI-enabled sensors, mobile technology, and facial recognition will intensify domestic surveillance.¹⁶¹ China is already utilizing surveillance technology to strengthen its authoritarian rule and has exported its policing technology to at least 80 countries.¹⁶² Observers allege that the export of Chinese surveillance technology will allow China to collate vast data to refine AI systems, improving domestic identification and tracking of dissidents.¹⁶³

Chinese companies, like Huawei, are reportedly collaborating with authorities to export its “safe city” solutions, designed to provide local authorities with surveillance tools such as facial recognition technology. Despite bans on Huawei 5G networks in several countries, Huawei has been actively selling “safe city” solutions globally, including the deployment of thousands of AI-powered cameras in Kenya and Serbia.¹⁶⁴ In the US, many view Huawei as an extension of the Chinese Communist Party (CCP), given the party’s considerable influence over Chinese private companies through heavy regulation. As such, the US views Huawei’s provision of AI infrastructure as another means for China to export its authoritarian model and fears such tools contain backdoors that allow the CCP to collate massive amounts of data to refine their own AI systems, attain access to critical infrastructure, and scale up espionage efforts.¹⁶⁵

b. Erroneous Surveillance by the Private Sector

Video surveillance and facial recognition systems are often developed by private sector companies who then partner with government agencies to collect and process personal information. In recent years, companies such as Clearview AI, Vigilant Solutions, and ODIN Intelligence, have been subject to concerns over the accuracy of their facial recognition algorithms and the diversity of their training dataset.¹⁶⁶ There is a general lack of transparency across the industry, with companies not legally required to allow third-party audits of their algorithms, and many do not or selectively publish their processes and results.¹⁶⁷ This lack of transparency raises concerns about private sector companies providing facial recognition

technology to government partners, particularly law enforcement agencies, with little to no due diligence. In the case of Clearview AI, the company broke Canadian law when it scraped the internet for 3 billion photos of people, created biometric identifiers from those photos, and then sold its facial recognition tool to law enforcement across Canada.¹⁶⁸ The example underscores how AI significantly increases the capacity to collect data on individuals. Data scraping photos on the internet allowed Clearview AI to build a database of approximately 10 billion photos, compared to the FBI which has 640 million photos.¹⁶⁹

c. Military Actions and AI Race

The combination of potential AI system failures and strategic pressures to integrate new technologies into military operations could potentially lead to an escalation in a crisis or conflict.¹⁷⁰ Experts suggest that offensive operations that incorporate AI or interfere with an adversary's AI systems could result in unforeseen system failures and cascading effects, triggering accidental escalation.¹⁷¹ Similarly, AI systems that are insecure, inadequately trained, or applied to unsuitable tasks could inject incorrect information into decision-making processes, inadvertently leading to escalation.¹⁷² The discovery of a compromise in an AI system could create uncertainties about the reliability of critical capabilities, potentially influencing decision-making toward deliberate escalation if conflict appears imminent.¹⁷³

These scenarios highlight the novel threats posed by AI advancements and their use on the battlefield. States seek to integrate AI systems to improve military decision-making, reduce uncertainty, and gain a better understanding of adversaries' intentions and capabilities. However, in relying on AI, states introduce a new source of uncertainty which may be prone to data poisoning attacks, and intentional or unintentional failures.

Furthermore, in providing AI tools for military operations, such as in Ukraine, tech companies hold outsized power as independent actors and raise questions as to the abidance of legal or regulatory rules and norms in pursuit of battlefield innovations.¹⁷⁴ The current war in Ukraine has been described as a "super lab of innovation" given the opportunities for tech companies to battle-test their AI systems.¹⁷⁵ Experts warn that the proliferation of such tools could risk falling into the hands of adversaries during conflict.¹⁷⁶ While most companies operating in Ukraine maintain that they are providing AI solutions in line with US national security goals, some question the sustainability of this arrangement and contingencies for ensuring these tools do not fall into the wrong hands.¹⁷⁷

Ethical concerns have also been raised regarding reliance on AI targeting in conflict zones. Commentators have warned that IDF's reliance on such tools could lead to "automation bias" whereby human operators are inclined to accept machine-generated recommendations, even in scenarios where humans would have reached different conclusions.¹⁷⁸

d. Terrorism and Bioterrorism

There are prospects of non-state actors leveraging AI for terrorism and even bioterrorism. The use of AI tools by TVEs has largely been used to propagate fake content, boost recruitment, and circumvent current moderation efforts. However, increasing attention has

turned to future use cases of generative AI by TVEs, such as enhancing operational planning which will pose more pressing risks to national security.

Researchers have studied the ability of TVEs to gather sensitive information from chatbots that could potentially be leveraged in terrorist attacks. Studies conducted in the first half of 2023 on Bing Chat and ChatGPT revealed that the systems could offer precise step-by-step instructions to significantly enhance terrorist operations.¹⁷⁹ These instructions included guidance on how to effectively remove online traces and information on which types of software may mitigate the risk of detection by law enforcement.¹⁸⁰ The chatbots also assisted in identifying cross-jurisdictional problems in sharing information on internet users that TVEs could exploit or even help generate simple scripts to remove data tracking features of operating systems.¹⁸¹ Both ChatGPT and Bing Chat also provided information on avoiding content takedowns and instructions on using the Ethereum Name System (ENS) that the Islamic State has exploited.¹⁸² Thus far, however, mainstream AI models have not provided more advanced information for operational planning, such as that shared in terrorist manuals.

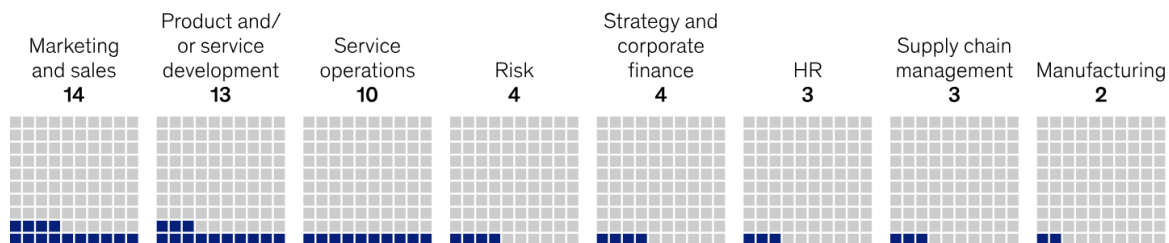
In terms of bioterrorism, the risks vary.¹⁸³ Scientifically naïve users, including some threat actors, may use chatbots to aid the information-gathering process, while scientifically knowledgeable users may be more likely to use chatbots to speed up routine tasks like gathering scientific literature and providing math or operational advice. Further, chatbots are useful for providing information on existing pathogens or toxins since they are trained on information that already exists. However, chatbots could also contribute to the development of novel pathogens by generating ideas. A malicious actor might prompt a chatbot to list immune targets or genes that contribute to pathogenicity and use this information to suggest potential risk-enhancing modifications.

F. Business Risks Due to Misuse of Generative AI

1. Overview

As the technical capabilities of AI systems have improved rapidly, businesses, governments, and other organizations have increasingly deployed AI tools, predominantly generative AI. Generative AI capabilities integrated into businesses include robotic process automation (39%), computer vision (34%), NL text understanding (33%), and virtual agents (33%).¹⁸⁴ The most common use cases include preparing sales and marketing strategies, generating code and content, developing software for product and service deployment, and conducting service operations. The figure below illustrates the main uses of generative AI by function. The survey was conducted across industries—including business, legal, retail, energy, financial, healthcare, technology, media, telecommunications, professional, and more—in organizations of varying sizes:

Share of respondents reporting that their organization is regularly using generative AI in given function, %¹



Most regularly reported generative AI use cases within function, % of respondents

Marketing and sales	Product and/or service development	Service operations
Crafting first drafts of text documents 9	Identifying trends in customer needs 7	Use of chatbots (eg, for customer service) 6
Personalized marketing 8	Drafting technical documents 5	Forecasting service trends or anomalies 5
Summarizing text documents 8	Creating new product designs 4	Creating first drafts of documents 5

¹Questions were asked of respondents who said their organizations have adopted AI in at least 1 business function. The data shown were rebased to represent all respondents.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

Figure 6: McKinsey Global Survey results on respondents' main uses of generative AI by function.¹⁸⁵

However, the increasing adoption of generative AI tools is not without inherent risks for businesses. According to McKinsey's Global Survey, few companies appear prepared for the widespread adoption of generative AI and the business risks these tools might entail.¹⁸⁶ Regarding specific risks of generative AI adoption, few respondents said their companies were mitigating inaccuracy in generative AI tools, the most cited risk.¹⁸⁷ The figure below details the generative AI-related risks businesses consider relevant and are working to mitigate.

Generative AI-related risks that organizations consider relevant and are working to mitigate,
% of respondents¹



¹Asked only of respondents whose organizations have adopted AI in at least 1 function. For both risks considered relevant and risks mitigated, n = 913.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

Figure 7: Generative AI-related risks that organizations consider relevant and are working to mitigate.¹⁸⁸

Although the above survey was conducted in 2023, the results are telling and still reflective of the overall lack of AI-related regulations to mitigate misuse. However, some progress has been made on official sanctioning or oversight of AI use in businesses: while Generative AI tools have been officially adopted by some businesses, others—including Apple, JP Morgan Chase, Citigroup, Deutsche Bank, Wells Fargo, and Verizon—have banned or restricted how employees can use AI platforms like ChatGPT.

2. AI-Related Threats

For businesses and the private sector, the AI-related threats are generally three-fold: AI-enabled dissemination and generation of vulnerable code, legal risks or challenges due to the adoption of advancing AI technology, and related insider threats.

a. Vulnerable Code Generation and Dissemination

AI code assistants using LLMs, such as GitHub Copilot or Amazon CodeWhisperer, have emerged as powerful tools to assist software developers in generating code much faster than before. While these tools can make developers more productive, businesses could face issues with the potential for automated code generation to introduce unintended vulnerabilities, such as weak encryption, injection attacks, or unintended access points. Recent studies found that AI-generated code presents security issues, and one study found that programmers using AI tools wrote less secure code than those who did not.¹⁸⁹ The study also found that the use of AI tools created a false sense of security among developers.¹⁹⁰

Furthermore, due to the inherent risk of hallucinations and biases in AI systems, generated code may produce false outputs that are difficult to detect and have negative impacts on

business functions such as reduced profitability.^{vi}

The use of AI in software development might also lead to the unintentional dissemination of source code which contains privacy risks. In one recent example, an engineer tried to send a source code snippet to ChatGPT that inadvertently included an API key that provided access to online services or applications.¹⁹¹ Though the issue was detected, this issue raises questions for software developers uploading code with access to sensitive information.

Currently, not all companies utilizing AI for business applications have sufficient security systems to detect, block, or remediate unintended data leakage although such leakage is a significant business risk.¹⁹² Another case includes unintentional data leakage at Samsung after engineers accidentally leaked internal source code to ChatGPT.¹⁹³ The input contained the source code of software responsible for the company's semiconductor equipment.¹⁹⁴ Due to concerns that data entered into platforms like ChatGPT cannot be retrieved or deleted, potentially disclosing intellectual property to unauthorized users, Samsung banned employee use of generative AI tools.

b. Legal Risks and Insider Threats

The widespread adoption of generative AI raises questions about intellectual property, legal risks, and compliance with new regulations. McKinsey's survey revealed that just 21% of respondents had established policies governing employees' use of generative AI tools in their work.¹⁹⁵ Companies adopting generative AI may face the following legal risks and challenges:

- **Exposure of sensitive data:** Employees risk inadvertently exposing confidential trade secrets and sensitive data such as Personal Identifiable Information (PII) by inputting data into generative AI tools.¹⁹⁶
- **Business liability for deliberate IP infringement:** Use of unlicensed material in training data can create unauthorized derivative works beyond fair use.¹⁹⁷
- **Accountability for mistakes caused by AI:** It is unclear who is responsible if an employee or business makes a mistake due to AI. Failure to properly handle and safeguard data can result in fines and reputational damage.
- **Code ownership:** If a business uses an existing AI model to write an application, there are questions over whether the original programmer still has ownership. In the US, the application of copyright laws to AI-generated code is currently unclear.

Despite common principles guiding AI regulation, the actual implementation and specific wording differ by regulator and region. AI regulation is still relatively new and subject to frequent updates, posing a challenge for long-term AI strategies in compliance with regulations. Differing AI regulations across regions can cause risks and confusion of non-compliance for developers of AI tools and other multinational companies using them. The

^{vi} A chief architect shared in his interview that hallucinations may be the biggest AI-induced business risk. In ML, a similar problem called a false positive occurred, and no concrete solutions were found.

following table outlines some of the key regulations and frameworks influencing businesses developing and deploying generative AI tools. Businesses and enterprises can base their own internal policies on the following frameworks and look to the regulations below as guidance in AI tool implementation and compliance.

Table 5: Overview of recent AI regulation impacting businesses.¹⁹⁸

Regulation	Implementation date	Description
European Union (EU) proposed AI Act	Early 2025 (approved in March 2024)	<ul style="list-style-type: none"> Stringent rules governing high-risk AI systems, transparency, and data governance measures. The latest draft bans the bulk scraping of facial images to build databases, social scoring, and emotion recognition in the workplace. Financial penalties for non-compliance, of up to 7% of annual global revenues. Though the act’s jurisdiction is limited to the EU, it will have extraterritorial impacts given its applicability to all entities with operations in the EU. Prior to the law coming into effect, the EU is asking companies to voluntarily commit to adhering to key parts of the Act by signing an AI Pact.
Biden Administration’s Executive Order (EO) on AI	October 2023	<ul style="list-style-type: none"> Establishes standards, tools, and tests to ensure the safety, security, and trustworthiness of AI systems. Mandates that developers of the most advanced AI systems disclose their safety test outcomes and other vital information to the US government.
NIST AI Risk Management Framework (AI RMF)	January 2023	<ul style="list-style-type: none"> Designed for voluntary use. Aimed at enhancing the capability to integrate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.
UK’s Principles-based Regulatory Framework	March 2023	<ul style="list-style-type: none"> Outlines “proportionate” rules for different sectors’ use of AI. Provides cross-sector principles including safety, security, transparency, fairness accountability, and governance. New standards to support regulators such as investing in the AI Standards Hub.
G7 Hiroshima Process	October 2023	<ul style="list-style-type: none"> The Hiroshima AI Process Comprehensive Policy Framework was established, including guiding principles and code of conduct aimed at promoting safe, secure, and trustworthy AI systems.

The use of generative AI by businesses also introduces enhanced insider risks—both intentional and unintentional—including new mechanisms for the deliberate leaking or theft of sensitive information, as well as novel risks including data poisoning attacks which can create false outputs (as discussed in Section 4.B.)

Experts are concerned that employees may not be adhering to generative AI policies in place, leaving the potential for intentional or unintentional data leakage.¹⁹⁹ One study found that over half of generative AI inputs in a sample contained sensitive or personally identifiable information.²⁰⁰ Another survey on data exposure reports that while most companies (99%)

have data protection solutions in place, these solutions are not mitigating data loss from insiders, particularly as a result of generative AI tools.²⁰¹ The insider threat raises concerns about the swift advancement and adoption of generative AI technologies outpacing organizations' efforts to update security policies and train employees on data exposure risks.

Another concern is the theft of sensitive information related to an organization's AI systems or usage. In early March 2024, a former software engineer at Google was charged with stealing AI trade secrets while secretly working with two companies based in China.²⁰² As AI becomes an increasingly contested space, US government officials have raised concerns about how foreign adversaries could harness AI technologies to negatively impact the US.²⁰³

c. Insufficient Oversight and Testing Procedures

Generative AI tools have changed traditional software development processes, which involve human oversight at each stage to ensure that each line of code adheres to established security protocols. With AI-generated code, oversight becomes increasingly difficult due to the increased speed of generation and human reliance and dependency on AI as a superior form of technology. If undetected, vulnerabilities can expose sensitive customer or proprietary data, risking a large-scale security breach.

The issue of benchmarking^{vii} highlights the challenges faced by businesses adopting generative AI tools, particularly those developing their own chatbots or virtual assistants. The process of benchmarking, while essential for evaluating the performance of AI models, may inadvertently introduce biases based on the selection of datasets and evaluation metrics. Thus, the fairness of comparisons between different systems may not be reliable. Benchmarking also requires significant resources, including extensive computational power and substantial time investments, which may pose challenges for organizations.

Currently, AI models are not judged according to unified standards, though there are increasing attempts to do so. A common benchmarking method is the massive multitask language understanding (MMLU) in which there is a bank of questions and answers to test an AI model on a wide range of tasks. Benchmarking gets complex, and a weakness of the test is that an LLM can be taught to beat the test. Thus, the benchmarks do not accurately assess how capable language models are at applying the knowledge they learn across different domains. As AI development rapidly progresses, benchmarking efforts also risk being outpaced by the evolving capabilities of these systems.²⁰⁴

Another popular approach is the head-to-head comparisons model. A human user interacts with two models at once, asks the same question, receives the two responses, and rates which response is better. This approach is an improvement because the model can't be taught to beat the test, but it is also a weak approach because it is dependent on the human user—complex, programming, or reasoning tasks can't be tested with this approach.²⁰⁵ Further,

^{vii} In the context of AI, benchmarking refers to “the process of comparing the performance of different AI models or systems using a predefined set of metrics, which enables organizations and researchers to determine the most effective and efficient approaches. Benchmarking plays a key role in refining AI algorithms, improving accuracy, and enhancing the overall functionality of AI systems.” Taken from: Lark Editorial Team, ‘Benchmarking’.

research has found that language models that tend to perform better on fairness benchmarks have worse gender bias.²⁰⁶

5. Risks and Impacts

Businesses adopting generative AI tools face various risks such as operation disruptions, financial losses, reputational damage, and regulatory noncompliance. The table below outlines the potential risks of businesses' use and misuse of AI systems and their impact.

Table 6: Summary of Risks Facing Businesses Adopting Generative AI Systems.

Risk	Description	Impact
Use of pre-trained models that are used directly or tailored to new datasets	<ul style="list-style-type: none"> External attacks on the tailored models can compromise the model, causing it to generate incorrect outputs or leak data.²⁰⁷ 	Operational Disruption, Reputational Damage, Financial Losses, Regulatory Noncompliance
Inadvertent data leakage and theft of trade secrets	<ul style="list-style-type: none"> Security policies for employees' use of generative AI are typically implemented on a domain-by-domain basis, posing challenges for implementing security policies on an overall domain.²⁰⁸ 	Operational Disruption, Reputational Damage, Regulatory Noncompliance
Financial firms rely on generative AI for risk modeling and financial advice	<ul style="list-style-type: none"> Potential data poisoning attacks compromising risk assessment models or providing inaccurate advice to shareholders^{viii} 	Operational Disruption, Reputational Damage, Financial Losses
Inherent vulnerabilities in AI systems (biases, incomplete testing)	<ul style="list-style-type: none"> Undetected flaws resulting in false outputs or user data exposure. Growing reliance on existing platforms like ChatGPT for business increases this risk. 	Reputational Damage, Financial Losses
Companies using generative AI customer service tools	<ul style="list-style-type: none"> Risk of violating consumer protection laws through poor implementation of generative AI tools. Chatbots producing biased language. 	Reputational Damage, Regulatory Noncompliance
Insufficient or unenforced generative AI-related policies	<ul style="list-style-type: none"> Organizations delaying action until regulations are finalized. Lack of prioritization in AI governance and organizational models could lead to the urgent need for remediation later due to regulatory changes, data breaches, or cyberattacks. 	Regulatory Non-Compliance, Operational Disruption

^{viii} Financial firms are beginning to adopt AI tools into its services and has already began using AI tools to develop credit risk models, assessments, and reports. Source: <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-generative-ai-can-help-banks-manage-risk-and-compliance>

V. Analysis and Key Findings

The following section presents analysis and key findings from the literature review as well as interviews with subject matter experts around cyber risks created by AI adoption and mitigation techniques.

1. Analysis of Research

Generative AI has broken out into the public consciousness, intensifying existing challenges and unveiling new risks for individuals, businesses, and society more broadly.

As industry continues to drive AI technology evolution, widespread adoption will be seen across society and especially in the technology sector. Individuals will be confronted with heightened privacy concerns to increasing deception due to AI-enhanced social engineering attacks. Businesses will also navigate various risks from the concentration of sensitive information into AI databases, heightening the potential for internal and external attacks, misapplication of AI systems, and regulatory noncompliance. These developments will have wide-reaching societal impacts, including AI-driven disinformation campaigns, pervasive ethical dilemmas, and enhanced cybersecurity challenges from the increasing use of AI tools by all sectors.

Generative AI acts as a force multiplier for cyberattacks, advancing existing TTPs.

The democratization of AI advances the speed and scale of traditional cyberattacks, potentially surpassing defenders' ability to adapt and respond effectively. The ability to harness AI capabilities will differ among different threat actors, with more sophisticated uses of AI enhancements limited to threat actors with significant resources and expertise. Over the next two years, the threat will arise from the evolution and enhancement of existing TTPs.²⁰⁹ Overall, threat actors are accelerating their business such as spear phishing campaigns with AI tools.

The democratization of AI has led to shifts in the cybercrime landscape.

Ransomware actors and other threat actors are already leveraging AI to enhance the efficiency and effectiveness of various cyber operations. However, AI lowers the barrier to entry for cybercriminals, allowing for personalized phishing campaigns, enhanced information-gathering, and more sophisticated malware generation. This trend follows growing concern that malicious AI models and jailbreaks are driving an exponential growth in phishing, given the speed at which AI allows cybercriminals to launch sophisticated attacks.

Generative AI tools are reshaping the landscape of disinformation.

Authoritarian governments are using AI as force multipliers for censorship, controlling tools, and platforms to manipulate information spread. As generative AI tools become more accessible, governments worldwide are likely to reinforce existing information controls, posing challenges to the free flow of information. The proliferation of deepfake videos exacerbates the normalization of disinformation, eroding trust in institutions and democratic processes over time. Governments and social media platforms face the challenge of countering falsified information while preserving principles of free speech and civil liberties.

Generative AI enables new types of cybercrime presenting challenges to mitigation efforts.

Instances of deepfake pornography have surged, driven by accessible AI image diffusion models. The promotion and sale of these services on social media platforms are raising concerns about online harm, including targeted harassment and extortion.

Inherent AI biases pose a significant threat to social infrastructure AI systems.

Biases inherent in training data are particularly concerning as failures can reinforce stereotypes and perpetuate discriminatory narratives. These AI vulnerabilities may lead to biased content generation or decision-making, raising ethical concerns regarding the perpetuation of existing discrimination.

Data poisoning attacks can alter decision-making processes, with severe consequences for critical infrastructure.

In the future, data poisoning threatens the reliability of critical AI systems, particularly in areas like the military, healthcare, criminal justice, housing, employment, and autonomous vehicles. Such risks could result in false outputs that inadvertently escalate crises, perpetuate biases, or lead to physical harm. Protecting AI tools against data poisoning remains a key concern to ensure the reliability and safety of critical systems.

The proliferation of AI tools is expected to intensify global espionage and mass surveillance.

Generative AI tools provide the ability to collect and synthesize data on a much larger scale than before, posing a serious national security risk. AI tools may allow authoritarian states to bolster their rule through advanced video surveillance technologies that facilitate the monitoring and tracking of individuals in public spaces, offering governments comprehensive profiles based on online activities and physical movements.

The integration of AI systems into military operations introduces novel risks, including implications for the use of AI for military decision-making.

Insecure or inadequately trained AI systems used in military decision-making may inject incorrect information into critical processes. Governments are integrating AI systems to enhance military decision-making and gain insights into adversaries' intentions and capabilities. However, this reliance introduces a new level of uncertainty, making these systems vulnerable to data poisoning attacks and both intentional and unintentional failures.

Non-state actors are increasingly leveraging AI tools to spread terrorist propaganda.

Terrorist Violent Extremists (TVEs) are increasingly exploring the use of generative images and videos to create synthetic content and evade content moderation efforts. More advanced uses, such as for operational planning, have not yet been seen but may pose a threat in the future as TVEs adapt to emerging technologies.

The inherent vulnerabilities of AI systems expose organizations to risks of reputational damage and regulatory noncompliance.

Undetected vulnerabilities—such as the risk of hallucinations or poor training/foundational data—can lead to the generation of false information or expose user data, especially as organizations increasingly rely on AI platforms like ChatGPT for critical business functions. Organizations face the challenge of ensuring compliance with evolving regulatory landscapes, creating business, legal, and reputational risks.

2. Key Findings from Expert Interviews

Ten expert interviews with individuals of various backgrounds—public and private sectors, engineering, legal, cybersecurity risk, academia, and more—were conducted to further enrich the analysis of this report. While the interview inserts can be found in section 6 of this report, this section summarizes the key findings.

Expert opinions vary on what the greatest AI-enabled business or cybersecurity risk is.

The risks that were mentioned include AI hallucinations and the lack of a solution, AI models regurgitating sensitive information, overdependency on AI, the force multiplying effect of LLMs and the lowering of the barrier to entry for novice hackers, uncertainty around AI opportunities and risks, and AI models lacking model “robustness” that fails to account for edge cases. However, there is more of an agreement regarding the most immediate and obvious AI-enabled cyberthreat: AI-enabled social engineering/phishing and disinformation campaigns. There is also a disagreement on what the greatest threat of LLMs is: some experts believe that the business risk due to insecure code generation by LLMs is understated while others worry about the potential generation of stealthy malware by LLMs.

The democratization of AI has been the leading cause of significant increase in AI usage.

AI technology has been slowly evolving for years, and the science and engineering community has been monitoring the progress. AI seems to have developed rapidly in the last couple of years because companies have been able to make AI accessible, understandable, and usable to the everyday user (i.e., Chat GPT gives the average person a way to interact with AI and use AI)

There is no one sector that is the most vulnerable now due to AI evolution.

All sectors have been victims of cyberattacks with governments increasingly concerned with cyberattacks against critical infrastructure for the widespread impact of such attacks. AI is a tool that enhances cyberattacks and business risks, and all sectors that use digital systems will experience enhanced risks.

AI framework and regulations are emerging.

Experts agree that the notable AI regulations thus far have been US President Biden’s Executive Order 14110 on AI that framed the issue, the European Union’s AI Act, and the US Office of Science and Technology Policy (OSTP)’s [AI Bill of Rights](#). Experts also highlighted that companies are entering into a voluntary commitment to a safer and more secure development of AI technology such as the Content Authenticity Initiative and the Coalition for Content Provenance and Authenticity. Some experts believe that eventually, regulations in the AI space will be distilled down to a couple of general best practices or regulations.

Combating AI-enabled disinformation will require efforts at the individual, technological, organizational, political, and international levels.

AI has been trained to be convincing, not necessarily correct, and can help reinforce extremist ideas. AI-enabled disinformation will impact the elections in 2024 even though there may not be a disastrous outcome or impact as feared by the public. Thus, ways to combat disinformation have been heavily discussed. At the individual level, public awareness of AI-generated fake content is the key to combating disinformation. At the technological level,

there are emerging ways to distinguish AI-generated content from the real by encrypting real content—German company Leica created a camera that encrypts real images. At the organizational and political level, irresponsible use of AI—especially in the news and media sector—will have to be regulated. Adobe is also leading an initiative to get the industry to mark AI-generated content. Having regulations on requiring AI-generated content to be marked, however, has sparked conversations on its utility and enforceability. Internationally, like-minded states need to create and abide by ethical standards aimed at combatting disinformation.

State-sponsored actors are already using AI tools.

Reports and experts alike have announced that state-sponsored threat actors are already using AI tools to help espionage and spear phishing attacks as well as help research vulnerabilities. Australia is currently undergoing its version of the US Shields Up campaign in preparation for a potential Taiwan Strait conflict and is preparing for Chinese actors to use AI-enabled disruptive attacks. Furthermore, state-sponsored actors have the resources to invest in the R&D of ML or reinforcement learning to conduct cyber operations. The use of AI tools will expand as R&D continues. Thus, governments and companies need to invest in AI now to test AI agents and algorithms in testing environments with guardrails and begin developing AI-enabled defenses.

AI technology continues to evolve rapidly.

Benchmarking for LLMs has been difficult because the benchmarking methods get beaten and outdated too quickly against rapidly evolving LLMs. On the other hand, specific security solutions such as AI firewalls or ways to prevent the AI model from seeing PII are starting to develop.

3. AI Threat and Risk Chart

In concluding this report, the following table summarizes the current threat landscape due to AI and is based on a literature review and surveys. The table attempts to categorize the types of threats arising from the increasing adoption of AI, as well as the impacted sectors or entities, and estimate the timeline of risk and threat level.

Table 7: Summary of AI Threats and Risk Chart.

Threat	Risk	Impacted sector/entity/etc.	Timeline ^{ix}	Impact
AI-enhanced traditional cyberattacks	Force multiplier for disruptive attacks	All sectors but critical infrastructure may be impacted greatly	Medium-term	High
	Increased capabilities, sophistication, and efficiency of cybercriminals in ransomware and cryptocurrency-related cyberattacks; lowered barrier to entry	Individuals and industries, especially ransomware-prone industries such as health care, financial, and hospitality sectors	Medium term	High
	Lowered barrier to entry for social engineering; increased efficiency and speed in spear phishing	Individuals, industries, governments, academia, news organizations, critical infrastructure	Immediate	High
AI-enabled disinformation	Domestic Disinformation: increased censorship, targeting of vulnerable groups, spread of authoritarian digital norms	Particularly individuals and minorities in authoritarian nations, democracy, freedom of speech	Immediate	Medium
	State-sponsored disinformation campaigns: polarization of societies, erosion of trust in institutions, degrading of democracy	Individuals, democratic governments, electoral process Democratic	Immediate	Medium
	Promotion of crime and discrimination: new class of crime such as deepfake pornography and stock market manipulation	Individuals, finance industry, black market, private sector widely	Medium term	Medium
	Election Obstruction: online censorship, disinformation	Individuals, freedom of speech, democratic nations, electoral process	Immediate	Medium-High
AI-Enabled disruption or maloperation of systems	Data poisoning: false outputs leading to bad decision-making, discrimination, disruption	Critical infrastructure, social infrastructure, justice system, others	Medium term	High
	Inherent biases and vulnerabilities: reinforce stereotypes, biased content generation and decision-making	Individuals, businesses, governments	Immediate	Medium
	Intentional and unintentional failures: operational disruption	Critical infrastructure, social infrastructure,	Immediate	Medium-High

^{ix} Immediate refers to happening currently or in the next couple of years. Medium-term refers to the next 3-5 years. Long-term refers to the next 5-10 years.

	and false outputs	justice system, multiple industries		
AI-enabled national security threats	Military applications: potential autonomous weapon systems, military decision making leading to ethical concerns	Defense sector, governments	Long term ²¹⁰	High
	AI race: deployment of AI systems with unproven reliability, risk of escalation	Governments, defense sector, industry	Long term	High
	Espionage and Mass Surveillance: higher scale and speed, erroneous uses by the private sector	Public and private sector, individuals, privacy	Medium term	Medium
	Terrorism: dissemination of propaganda, assist with terrorist plans	Social media companies, individuals, governments	Medium term	Low
	Bioterrorism: development of novel pathogens, efficient information gathering	Individuals, healthcare, and pharmaceutical sectors	Long term	Low
Business risks due to misuse of generative AI	Vulnerable code generation and dissemination (can be due to insufficient oversight and testing): data leakage, reputational damage, regulatory noncompliance, financial losses, operational disruption	Businesses, consumers, employees, privacy	Immediate	Medium
	Legal risks and insider threats: data leakage, trade secret theft, noncompliance, financial penalties	Legal system, privacy, businesses, individuals	Immediate	Medium

6. Expert Interviews

1. Chief Architect 1

How is AI increasing the risk and impacts of traditional cyberattacks such as ransomware, social engineering, and disruptive attacks?

I don't think that attackers are fully exploring the full potential of AI yet. They are just starting to use AI tools, and there was a recent Microsoft article on how specific actor groups are using generative AI. Attackers can use fairly simple techniques to get into target systems right now, so there is no reason to fully utilize AI yet: the defenses have not caught up to the point where adversaries can't make money or achieve their objective without using AI. As defenses get better, attacks will start using this underutilized toolset—generative AI will allow for a very high scale version of what humans can do in a cyberattack or campaign. Adversaries are using a couple of tools and sending out phishing emails right now, but employing AI will be the equivalent of thousands of attackers working concurrently. A thousand concurrent workers trying a bunch of tactics to see what works would be going beyond what the human is limited to. So, AI's risk and impact are about scaling to refine human tradecraft and automating it.

How can businesses and organizations improve their defenses against potential AI-enabled attacks?

A key point about AI is that the stronger the weapon you make, you can simultaneously make a defense system for that exact weapon; this doesn't necessarily work in physical warfare. With AI, there are techniques that developers can take from adversarial systems and build a defensive system that is just as good. I expect to see a bigger investment in simulations and systems to test their AI defense systems. Organizations can build an AI weapon and use it to primarily train and improve the defense system. I have not seen this on a large scale just yet because it is a hard problem. It is one thing to generate a phishing email with Chat GPT, but it is another thing to make it scalable and do it well. People have been thinking about AI businesses and product models for a long time, but entire SOC environments will have to fundamentally change because attack surfaces are only getting more complicated. There just aren't companies making defense systems for this completely new environment yet. There are companies like Horizon 3AI doing AI-driven red teaming, but we have not seen anything mind-blowing yet.

How and why has generative AI developed so quickly in the last two years?

From a technical standpoint, language models and even LLMs have existed for a while. The real breakthrough was not one thing like GPT. GPT founded its development on multiple things, including a research paper from Google. The real breakthrough was from people figuring out how to scale the model to the scale of the internet. Earlier, we didn't have the computability or the ability to collect the amount of necessary data to be able to learn on the collective intelligence on the entire web. Once the models were able to digest large data sets, we started seeing emergent behavior—behaviors from the system or model that we did not explicitly teach—such as generative AI's ability to reassemble code well. So, the scalability was one reason.

Another factor was making AI accessible, understandable, and available to the average user. It's interesting that GPT3 which was released like three years ago did 96% of what Chat GPT does now. The science and engineering communities were amazed, but the technology was confined to that community because it was an API-centric tool and made for developers. Chat GPT brought the technology to the everyday user, marketed it well, and reached a wide audience. Image generation was on a different track but underwent a similar transition at a similar time. For example, Mid Journey scaled taking all the images of the web to generate a new image and made the tool pragmatic.

What is the biggest business risk due to AI usage?

There is no known solution to solve hallucinations. In classical machine learning, we had this problem as well but called it a false positive. For some reason, intelligent models went off the rails in obvious cases, and we couldn't fix it. It's even more complicated now because the models are larger, more complicated, and harder to understand.

Another risk is information leaks. We already see examples of people crafting prompts and inputs that trigger the model to leak information. It's dangerous because often, an organization doesn't even know all the data going into these AI models. PII or even core IP gets regurgitated by models even though they aren't supposed to release that information—the knowledge is baked in the model and should not be shared.

Have you seen any AI specific security solutions or defenses arising in reaction to the security threat posed by AI usage?

We are already seeing companies and startups focus on making AI firewalls, for example, to block some data from getting leaked by AI systems and models. There's also work being done on how to stop PII from even entering a model or to prevent the model from ever seeing the PII—this would be work on the AI infrastructure side.

Do you think labeling AI-generated content is an effective method to prevent AI-enabled disinformation campaigns?

Deep fakes and generative AI generated personas will take off this year with the US presidential election. We don't really know what the complete attack surface looks like. I don't think that labeling AI-generated content is enforceable, and the inability to keep it consistent will create more of a problem because it creates a sense of false security. However, Facebook, for example, already marks content that it believes to be AI-generated. Facebook was naturally incentivized to do this because there is a business risk if the company's news feed is just full of fake content. These natural incentives in the system will help mark AI-generated content, but I don't see a great way to enforce this in a legal or policy framework.

2. Public and Private Sector Expert 1

What has changed the most in the last two years in terms of AI?

What has changed the most is that people are a lot more desensitized to AI to the point of AI becoming common and accessible to the point of a job hunter using AI to write cover letters and thank you notes and to the point of everyone talking about ChatGPT. People now see AI as a tool with practical applications—such as job hunting—and creative applications.

How has the development of AI changed the international power dynamics?

With the development of AI, the US has taken a globally dominant position in the discussion. People have and do talk about the China threat in terms of AI with exfiltrated US research and development, but when you look at who is creating the products and innovating AI technology, it's the US. In the marketplace, US-made AI products have taken over, and the US is leading the AI revolution.

What is an increasing concern regarding AI?

A common concern is AI governance. There are a lot of soft power discussions that are delaying the evolution of AI. Starting the latter half of 2023, the federal government has been trying to look for governance models that work across multiple fields in AI. NIST has taken a strong position in trying to create a risk management framework related to AI as well.

We are also at an interesting point now where we need to determine where the line between what a human does and what AI does is. If people are now using AI to write cover letters and thank you notes, does an employer even need to request a cover letter now? Is this a performance exercise that should stop?

Combining these two thoughts, there are a lot of discussions, talks, and white papers on what frameworks we should have and what AI controls should minimize. I think we will eventually settle on two or three broad frameworks, like the ones in the cybersecurity realm—we will need to resist the urge to think that one size will fit all. People in this space will come to realize that they need to adhere to or at least acknowledge the frameworks to become legitimate in the environment. There needs to be an ethical standard for AI use, and we will probably need to keep humans involved in the decision and evaluation points even if humans use AI to tee up the decisions and evaluations for discussion. For example, is using AI to determine which intelligence operations to run, how to minimize economic threat, and which technologies to invest in where we should be applying AI?

On the creative side, there have been evolving conversations and tensions with creators and those who are using AI to generate creation—such as lawsuits regarding what is included in an AI database to create imagery or video. There will eventually be limits and questions on what we use machines for versus what we use humans for. Again, drawing that line will be the next important phase.

Another concern is the media over-fixating on AI-generated distractions. Mal-intentioned actors will use AI to foment distraction, and are we as a population, and news media organizations, being responsible for what we choose to cover? This makes furthering knowledge and awareness important as well as ethical standards of and training for journalistic integrity, not just sensationalism, important.

Why are people and societies wary of AI evolution?

There is a huge resistance to leveraging AI technology because of fear: fear of people trained on older models and fear of having AI machines make certain determinations. I have heard before that AI is trained on models to be convincing, not correct. I think that is the crux of why we have resistance and the reason why we have people gravitating towards AI technology. AI is going to continue becoming positively integrated into our societies and systems, but in order to be more broadly integrated, we are going to need to have an expectation of AI being correct, not just convincing.

In reality, AI has not been making the sky come crashing down. Over the Pakistani election, a Pakistani leader who was imprisoned used an AI generated message to provide rousing political rhetoric to his constituency and to prevent a military overthrow or a compromised judiciary. Having proxies use AI to create political messaging led to a motivated voting populace despite military rule that imprisoned and disposed a leader before re-election. I think this makes for a great story. AI didn't break the system nor drastically shift the conversation/electoral outcome. I think two years ago, people were worried that AI would break the democratic process; this story shows that maybe, AI is just the new speechwriter.

How can we as a society push for the expectation of AI being correct and not just convincing? In other words, how do we combat disinformation?

With consumer demand and wider application of AI, I believe that the correctness of AI will be necessitated. We will always have a margin of error, but in some spaces, we don't have a lot of wiggle room. Calibrating what an acceptable margin of error is will be something that will be determined by different sectors.

There is no fool-proof way to detect disinformation although Adobe and other companies are already creating marking technology to flag the real content or to label and tag real content based on a technical threshold. Various companies are joining consortiums to commit to labeling things that are AI-generated. I think it will also depend on regulating the news media industry to have healthy and balanced reporting from trusted news sources. There is also a role for the K-12 education system. Especially after the COVID-19 pandemic, our students are taught about cyber bullying, cybersecurity, and password protection. We should add AI to such education to teach our future workforce when leveraging AI is okay and when it is not. It will be like applying the NIST framework for the education system and professional workforce.

3. Public and Private Sector Expert 2

How is AI increasing the risks and impacts of disruptive attacks?

Generative AI lowers the cost of entry for disruptive attacks. Script kiddies have easy access to the internet, and now there is a new volume of disruptive attacks. There is a higher level of aptitude since AI crowd sources understanding and best practices for penetration into networks. There is also an automated process with AI rapidly identifying vulnerabilities and directing hackers to the vulnerability. Before, there would be a vulnerability, and a company would have a few days to patch it before someone finds the vulnerability. Now, the vulnerability is found within seconds.

The impact of such disruptive attacks is much greater now as well. There is increased situational awareness and contextualization of networks—hackers know more about the network. For the defender or the network owner, it will come down to “can I manipulate an adversary into thinking that they know the true context or situation of the network”? AI enables this cat and mouse game in which defenders will have to create more noise in the network activity to create obfuscation and confuse the attacker: As much as hackers feel like they have better context and understanding of attack vectors, they are confined into weaknesses of current algorithms and network owners’ manipulations of the network.

On the other hand, we can now see prompt engineering as the first indications of AI-enhanced disruptive cyberattacks. We can build defense mechanisms against this indication and sell the mechanisms to better position defenses. There is a lot of experimentation and testing being done which creates footprints for us to use to identify APTs and understand their risks.

What are the greatest risks and threats posed by the (mis)use of AI systems in the defense sector?

Everyone has to get comfortable with the highly automated, but the fear of AI usage and error keeps the government supporting a “human in the loop”. For cyber and hypersonic defenses, the human in the loop is nonsensical, and in fact, the human gap is a vulnerability. We need to fully automate, lean forward, create AI defensive shields, and track and mitigate actors before they come into our domains. Of course, the attacker will have the same AI advantage, and this will be abused. It’s a matter of who can understand and control the AI systems and models as well as who gets the policy—such as the Department of Defense (DoD) policy about automated weapons system—behind it. We must explore the boundaries of AI to fully appreciate AI, and we are getting better at it. The volume of threat coming at us due to AI is already beyond human scale, so we need to work hard to implement AI systems as well while following the [DoD principles](#). It is not a time for fear, but a time for solid academic and professional tradecraft with guardrails.

How can we begin pushing this boundary around AI systems and be more aggressive with testing and innovation?

First, is to test AI with guardrails in a confined environment. If the AI agent works well in that environment, then the AI agent can be further tested or employed. We are reaching a level

of higher computation where we need AI watching AI watching AI in testing/experimentation, and this process will help build confidence. This is building and honing tradecraft responsibly.

We can also use simulation environments to put AI agents against other AI agents, use high fidelity metrics, and take out agents that are moving towards lower accuracy. There should be crosschecking at every step of the process and algorithm implementation, and computability allows us to do so while having minimum impact to performance.

We can also bring in outside algorithms to join these ensembles, and experimenters can vote on which algorithm performed the best under certain scenarios. Algorithms that are moving towards being confidently wrong or demonstrating a low accuracy mode, those algorithms will get removed and fenced for the rest of the engagement.

All these tests and simulations require investment into creating an environment that allows us to pit agents against agents and that produces good data that is translatable to the real world. We had a lot of investment in cyber ranges within the last few decades; we now need a lot of investment on AI for proving ground, and the government has an opportunity to invest in that. Then, there needs to be a grading criterion. The strength of having an ensemble of algorithms is to be able to gain the understanding of the weakest link in the algorithm. Then, we can set up guardrails there. However, to gain this understanding, we need investment; we need to convince the US government that AI is ready.

What are some potential risks in deploying AI to decision making or military operations?

Generative AI is designed to be confident even with flaws. We should build our trust in AI, but at what point does that trust become dependency. Once we become dependent on the machine's answers without challenging the answers, the value curve of AI begins to go downhill. We cannot give up all thinking because we deploy AI.

How are nation state actors using AI?

There was a Guardian article on how the big 4—North Korea, Iran, China, Russia—are using Generative AI to observe companies in various sectors like the telecommunication sector.

4. National Organization Executive 1

How has AI significantly evolved in the last two years? Which AI-enabled risks have emerged due to this?

Most significantly Generative AI has become mainstream which is opening new risks. ChatGPT has kicked off a new moment for public use and awareness of LLMs. The mainstreaming of generative AI tools happens at the same time as main social media platforms seem to disinvest from general safety safeguards, for instance by reducing the size of their Trust and Safety team (see for instance Twitter/X for the most acute example of that trend).

What do you see as the most promising way to combat disinformation? Please feel free to include examples of specific applications or initiatives.

Post 2017, a series of interventions and transparency measures have been tested and pioneered by platforms to tackle disinformation: they have largely been abandoned or disinvested and should be reignited. This includes, for instance, comprehensive databases of incidents of foreign interference detected and remediated by platforms.

What regulations do you see impacting the environment and reducing AI risks?

Content moderation focused regulations, such as the EU Digital Services Act, will have a great impact.

Have you seen any AI specific security solutions or defense arise in reaction to the new presence of AI threats?

Watermarking has taken the spotlight when discussing solutions, as evidenced by the platforms commitments taken this year during the Munich Security Conference (see <https://blogs.microsoft.com/on-the-issues/2024/02/16/ai-deepfakes-elections-munich-tech-accord/>). I personally think that the deepfake/watermarking focus as a key problem/solution space oversimplifies the types of threats we'll see as a result of the mainstreaming of generative AI and constrains the set of innovative solutions we'll need to properly tackle them.

How do you think AI systems assist or enhance terrorist activities?

Much ink has been spilled on how AI will enhance the capabilities of bad actors, and I'm grateful to scholars who are reframing the conversation to ensure we can think through the *marginal* risks of these capabilities (see for instance: <https://www.aisnakeoil.com/p/on-the-societal-impact-of-open-foundation>). With that in mind, bad actors are likely to benefit from the same economies of speed and scale than others when turning to new AI tools to enhance their workflows, but I haven't seen evidence (yet) that harms and risks would be radically different.

5. Financial Sector CSO/CISO 1

What are some of the biggest business risks due to recent AI evolution?

AI and decision-making process comes down to how well the model has been trained and what training data set has been incorporated. The biggest business risks are the edge cases where the decision or outcome may have not been the intended outcome/expectation. AI implementation scales up the way we do business and process at machine speed which leads to a larger impact. If AI systems derive the incorrect outcome, the risk to businesses and operations will be impactful. A human making a mistake versus a machine making a mistake: they lead to drastically different levels of impact. Human mistakes can be negligible while machine errors can be rapid and scale quickly. We need to think about the quality of and level of confidence in machines to deliver the correct output. Edge cases stress the decision making. I call this “model robustness” – how does the model behave with edge cases.^x

I don't think there is much of a human or labor risk since the workforce skill set that businesses look for will just change.

How are AI systems deployed in organizations and businesses? How will they be deployed?

AI chat bots have been deployed for a while. ChatGPT is newly available for mass consumption, but prior to ChatGPT, AI was embedded into chat services. However, because AI is now available for mass consumption, the realms of possibility are endless: AI models no longer need large amounts of training data since everyday users provide the data. In the next five years, Generative AI will start doing knowledge workers' work, but knowledge workers will have to train models. The more we use AI and input data, the more skillset we will need to manage, oversee, and train AI. AI deployment will automate low-level work, make organizations more efficient, and shift people's roles and responsibilities.

What are the greatest risks and threats posed by the (mis)use of AI systems?

Adversaries are using AI to identify vulnerabilities quicker and execute attacks quicker. Now, any digital system is more vulnerable, and this is not sector specific. Digital companies are struggling to keep up with sophisticated adversaries using AI: the script kiddies have now levelled up. From a business perspective, there are new supply chain risks. Supply chains were already vulnerable, but with AI, there is a risk of expanded and faster impact.

What regulations do you see impacting the environment and reducing AI risks?

I think the European Union's (EU) AI Act and the US Executive Order on AI will be the two leading regulations that will steer the rest of the globe. Australia is drafting its AI regulation already to parallel the two. India is also investing a lot of time into AI regulations. New Zealand has a regulatory working group started as well.

^x Edge cases refers to situations or instances that do not follow norms, patterns, or averages. A financial sector CSO/CISO defines model robustness as the ability of an AI model to produce proper outputs for unusual and extreme cases.

What part of the world is driving AI innovation?

The US feels like a prominent AI innovator, but there are pockets of innovation elsewhere. There is a lot of capability coming from Canada, and the University of Waterloo is recognized for its AI research. Australia is developing quickly, and so is the UK.

How does AI increase the impact of disinformation?

People don't validate information in the news and social media already which has been driving division within countries. AI will further enable the extremes since AI will see an extreme belief reflected in a campaign or content, train on it, and reinforce it. Fallacies are being continually supported which will cause a broader divide. Democracy will not fall apart in 2024, but many nations vote in 2024 (India, Russia, US, Korea, Taiwan). Thus, the geopolitical risk of a fall out is greater due to the sheer number of elections. People are going to vote for governments that are extreme, those extremists will have their own executive orders and agenda, and that will challenge democracies.

How are state-sponsored actors using AI?

In the immediate term, they are using AI for disruption or espionage. Disruption is easy, too. Australia is already doing something like the US Shields Up Campaign because it expects China to move on the One China Policy by 2027.

6. Cyber Security Company CEO 1

What are some of the biggest business risks due to the recent changes in AI technology?

I think the biggest business risk is uncertainty. Most executives don't understand AI nor its opportunities and consequences. AI is not sufficiently tested yet, and people are uneducated or poorly informed which lead to unpredictable behavior. People need to understand the opportunities and potential risks to make business decisions accordingly. We have to remember that the explosion in Generative AI is relatively new. People used to have access to applications that used AI, but now the common people have access to AI itself: admins, developers, office workers, and soccer moms all have access to AI for business and everyday tasks. This means that there is a whole new range of things that people can do, and people have not been able to make sense of the range yet.

How are AI systems deployed in businesses and organizations?

People have access to Generative AI like ChatGPT. Third parties use AI. Platforms reach out to and are embedded with AI. Companies are merging or partnering with AI companies. Developers are accessing AI through applications program interfaces (API). These areas are all the areas where we need guardrails on AI usage.

What are the greatest risks and threats posed by AI usage in businesses?

The greatest risk is from software developers accessing AI through APIs. Developers use AI-enabled API for routine work which requires input of data. This also means that the developers are spilling out some important information into the cloud, and other AI models/systems are now collecting and training on your data. A smaller risk is when workers just use AI for tedious tasks. For example, a board secretary uses ChatGPT to proofread third quarter earnings the day before the earnings report is released. The secretary has now exposed the document to the cloud and AI system which can lead to a premature release of the data.

What are potential guardrails/regulations on developers using AI-enabled APIs?

The first step is to understand the problem. OWASP (Open Web Application Security Project) and other development organizations are already coming up with best practices for developers using AI and LLMs. I believe that the guardrails start with best practices; then the vendors will come across and help make the practices into controls that mitigate the risk.

The Biden Executive Order is also good and lays out basic principles well. Executive Orders are effective when they raise the right principles and create a new framing of an issue. The Executive Order framed the problems of AI well, and it is operational in the sense that it is influencing discussions and startups. Regulations will be hard in the AI space because each example is so deep, rich, and complex. For example, is sending AI-embedded robots into Gaza to find victims a good idea? These questions cannot be answered so easily nor regulated. Thus, the best approach will be to have a strong set of principles that allow the AI user to be responsible for balancing the good and the bad.

One part of the Executive Order that I do not agree with is that American should know if something they are using is AI-generated. For example, if Walmart is using an excel spreadsheet to develop its earnings report, do we mark that document as “created by excel”? No, we do not. In the same way, AI is a tool, and the combination of human and AI generation is the right way to go. To say that we must mark all AI-generated content implies that AI is worse than a human or the AI user. AI hallucinates and has biases but compared to humans, AI can be more correct. We keep comparing AI to perfection and fear that AI falls short of perfection; however, we should be comparing AI to humans and use it as a tool. Furthermore, I do not think that marking AI-generated content will help people agree on issues as the US is polarized already. Labeling AI is going to be meaningless.

The Office of Science and Technology Policy (OSTP) also developed an [AI Bill of Rights](#) which is also good.

7. Cyber Security Company CEO 2

What is the most immediate national security threat you foresee due to AI-enabled attacks?

The immediate national security threat is not confined to one type of attack but is related to the force multiplying effect of LLMs. LLMs become a force multiplier that enables fewer people with less knowledge to do more badness. We and our sister nations are witnessing malicious actors primarily gain access through phishing, and LLMs allow perpetrators to create more realistic personas. LLMs bridge the language barrier, help perpetrators target victims, and gets precise with spelling, cultural nuances, and grammar. Without AI, understanding the cultural nuance and local/regional dialect would take much more time and labor. AI creates the propensity for victims to make that faithful click that allows perpetrators to get the initial access and to leverage that access to leap through networks and do lateral movement.

What are some likely state-sponsored AI threat scenarios?

With North Korea, since it has been trying to generate revenue for its weapons program's R&D as well as deployment, North Korea-related entities will use AI-assisted tradecraft for deeper and broader ransomware penetration. With China, since it has been targeting nations like the US and Japan that can pose potential threats in the future or that are entities of interest for IP theft, Chinese government individuals and contracted entities will use AI to enhance their tradecraft, accuracy, and efficiency to extract more information or to better preposition—it could preposition malware in a critical infrastructure system for a future disruption attack.

What do you see as the most effective mitigations to combat AI-based threats?

There are some basic mitigations such as using two-factor authentication, keeping apps updated, and making strong passwords. Everyday users and even governments don't like taking all the necessary steps for cybersecurity hygiene because it's a hassle and a nuisance. Taking these basic steps would push back on AI-enabled malicious actors from gaining credentialed access. With AI and supercomputing, guessing passwords is even easier. The weakest link is the one with human in the loop now because encryption is so strong, so basic cybersecurity hygiene comes down to where human engages the machine.

What do you see as the most promising way to combat disinformation?

Disinformation is an anxiety-inducing category because it creates fictitious synthetic personas by synthesizing real voices or images into a model. We already see falsehood being used in elections as we saw with China's attempts in the 2024 Taiwanese presidential election and the fake Biden robocall before the New Hampshire primary. So, last fall, like-minded nations met in the UK to create ethical standards on AI. But laws only help those who abide by the law, and autocratic nations will pick and choose which norms and standards to apply. Because of this and even more so, like-minded nations must come together and collaborate to create standards and abide by them. International norms on the criminal accountability piece of AI-generated content are still young, but in progress. In 2016, Russia used fake personas to

influence the 2016 US presidential election, and the US pushed backed. By 2018, the troll farm in St. Petersburg was dismantled.

8. National Security Expert 1

How is AI increasing the risks and impacts of common cyberattacks such as ransomware, social engineering, or disruptive attacks?

I think the most obvious increase in impact is with social engineering because AI has been allowing actors to imitate voices or audio and faces with images and videos. Imitating a specific human seems to be the biggest new risk, and generative AI enhances social engineering with audio, video, and image models. AI for face swapping is also very relevant, but it is a different branch. Social engineering attacks and use of disinformation, enabled by AI, to promote political ends are increasing risks. While we are going to have to continue monitoring all threats to models, Generative AI and LLMs will sharpen attacks. People will find creative ways to use multipurpose AI tools to make current threats more serious. I am sure that new threats that we can't even predict will also rise.

How has AI significantly evolved in the last two years?

Generative AI has been developing the fastest, and in particular, image generation and language generation (LLM). Having interacted with LLMs in the last two years, I have seen the types of responses from LLMs improve significantly. The quality of images has also increased to be more detailed, realistic, and varying. LLMs are also a problem because benchmarking methods can't keep up with the development of the models. Benchmarks get saturated and beaten quickly.

What are some common ways of benchmarking LLMs?

A common benchmarking method is the Massive Multitask Language Understanding (MMLU) benchmark. The MMLU has a question bank with template answers, and we test different LLMs with the questions to evaluate the answers. A weakness in this benchmarking is that you can teach the Large Language Model to the test which is not very demonstrative of the model's actual capabilities. Another popular approach is the head-to-head comparisons model. A human user interacts with two models at once, asks the same question, receives the two responses, and rates which response is better. This approach is an improvement because the model can't be taught to beat the test, but it is also a weak approach because it is dependent on the human user—complex, programming, or reasoning tasks can't be tested with this approach. Benchmarking continues to be an open challenge.

How threatening is AI-enabled disinformation, in your opinion?

Disinformation may not be as much of a problem as we expect. I don't typically believe that deep fakes will destroy the meaning of truth. Society quickly adapted to photoshop, and I do think it will adopt to deep fakes as well. Disinformation may be a minor problem in the 2024 US presidential election, and until we have one or two serious deep fake issues, disinformation may not be incredibly impactful.

What do you see as the most effective mitigation to combat AI-enabled disinformation, especially disinformation?

Awareness for sure. As people get used to AI, AI-enabled threats will get less effective, and especially so with social engineering.

Biden's Executive Order (E.O. 14110) on AI includes instructions for federal agencies to authenticate US government-produced content. Over 2023 Summer, seven leading AI companies (Amazon, Anthropic, Google, Inflection, Meta, Microsoft, OpenAI) made voluntary commitments to a safer and more secure development of AI technology. One development is watermarking fake audio and visual content.

There are also rising initiatives such as the Content Authenticity Initiative (CAI) which focus on systems to provide context and history for digital media. In particular, the Coalition for Content Provenance and Authenticity (C2PA), an initiative led by Adobe, is trying to address dis and misinformation by establishing technical standards.

As reliably detecting fake content becomes increasingly difficult, we will probably begin to have better systems for tagging real content as well. The most promising method right now is embedding something in the hardware to mark a real image. For example, a chip in the camera to encrypt real images with a mark. The camera company Leica developed such a camera/technology last year—Leica also implemented the C2PA Standard. Over the longer term, companies will figure out how to do this with microphones as well to encrypt audio files. As this kind of hardware technology is developed, a new mindset will have to come in place: they are fake unless proven real. The speed of adoption of this method and mindset will rely heavily on hardware companies' efforts to develop the technology.

What do you think is an increasing AI-enabled threat in the longer term.

None of the initiatives I mentioned apply to generative text right now. Generative text will be a more serious risk over the longer term. Most of LLM progress is the most meaningful progress to creating intelligent machines. Although speculative, I think a wider and deeper set of risks come from this cognitively capable AI systems. Also, generative text makes creating offensive cyber capabilities more accessible to the script kiddy and unsophisticated actors.

What is the most immediate national security threat you foresee due to AI-enabled cyberattacks?

Most concerning would be the democratization of cyber offense, allowing more people to act like sophisticated APTs and nation-state actors. Hackers will have their personalized AI assistants who are good programmers—this concern is mostly speculative at the moment, and there is no evidence of this situation yet.

State-sponsored actors are already very capable and sophisticated. AI generally enables larger scale, targeted, and high-quality attacks, which state-sponsored actors can typically already do. State-sponsored actors may be able to oversee more cyberattacks and campaigns due to AI though. State-sponsored actors already use AI to generate deep fakes like how the Russians

use deep fakes in the War against Ukraine. State-sponsored disinformation campaigns are the most likely and immediate national security threat, but it isn't anything new, severe, nor radical.

9. Cyber AI Research Expert 1

What are some of the biggest cybersecurity risks that businesses and private organizations face due to recent changes in AI technology?

First, for a technology company, there has been a lead in the past year or so since the release of chat GPT, where the LLMs have a fairly good capacity to write functional and good code. There has been an eruption of individuals, software developers, and security professionals leveraging code generating capabilities to write all sorts of code. As all outputs go, sometimes they hallucinate or give erroneous code, but the outputs are functional. Developers are not 100% relying on these tools but are using them on a semi-regular basis—AI-generated code is already folded into software supply chain. AI-generated code is also shared into open-source repositories, like GitHub. The problem is that while these codes are functional, the codes can be insecure—as in buggy or have errors like buffer overflow errors that can allow exploitation.

For big businesses like Amazon, Meta, or Microsoft, this is not much of a problem because those companies already have robust code review processes and in-house security engineers; however, this is a problem for small startups/businesses and individuals. The more AI-generated codes are functional, the more we psychologically trust the output and the AI-generated code. There are downstream risks, and new attack vectors that make the ecosystem more unsecure. As more unsecure codes get on open-source repositories and future LLMs do a web crawl/scrape, these unsecure codes will train next generation AI models, and then more unsecure outputs will be generated. Although there are no serious cases of this yet, Meta released a [publication](#) last year, that explored this issue: GPT4 has more insecure outputs even though the model is more functional. I believe that the risk of LLMs writing polymorphic code and malware that will enhance social engineering attacks are overhyped while this risk of AI-generated insecure code is underhyped.

How has AI increased the risks, impacts, and timelines of already common cyberattacks?

Ransomware attacks and social engineering attacks are probably the low hanging fruits that are most immediate with lower impacts. Even novices can use AI to conduct social engineering attacks for access and then exfiltrate data for a ransomware attack.

In considering disinformation campaigns and phishing attacks, the phenomenon is the same: use Generative AI to persuade a target to act—whether that is to click a link or to believe something. The consequences are not catastrophic, but AI can increase the volume of such attacks. Some early research suggests that LLM-generated phishing emails or disinformation word snippets are not yet as convincing as a trained human writing email or running influence operation; but there is a tradeoff between quality and efficiency. LLM-generated content is quite convincing, not as convincing as human-generated content, and quick to create. [Industry reports](#) have already highlighted that phishing emails skyrocketed by 1000% since the introduction of ChatGPT. Then, opportunistic cyberattacks like ransomware attacks will also increase in tandem with increasing phishing attacks.

What are the new categories of threat that will emerge due to AI?

AI will augment existing TTPs and will also create new categories of threat. Some examples include the following against AI companies and systems: adversarial machine learning—poisoning model data; extracting model data to train a copycat version; extracting private information behind the model; prompt injections; planting sleeper agents inside models that generate malicious codes under certain conditions.

How will state-sponsored actors use AI in their cyberattacks?

Nation state actors would first use AI to do vulnerability and zero-day research. Countries like China would stockpile those vulnerabilities while opportunistic actors like DPRK will use those vulnerabilities to scale up their cybercrime activity and make cash. LLMs have a lot of promise in scaling up certain cyber operation tasks that used to be manual, so state-sponsored actors can use AI to increase the volume of their cyber operations. When nations know that other actors are using AI in these ways, nations start doing it themselves, leading to AI arms race dynamic.

Also, not all AI is Generative AI. We have machine learning to detect anomalies or reinforcement learning—giving small rewards and punishments to a model to guide the model to learn a particular strategy within a confined rules-based system—to do cyber defense and maybe offense, too. State-sponsored actors are more resourced, so they can invest into these R&D intensive, high-end, sophisticated ways to apply AI to cyber operations. CYBERCOM and NSA are scrambling to get more AI research and applications going on as shown by DARPA's new AI Cyber Challenge.

10. Venture Capital Investor and Philanthropist 1

What has driven the evolution of AI in the last two years?

What has rapidly revolutionized AI is the speed of computing. [OpenAI uses 17,000 times more](#) electricity than the average American household, meaning there is something meaningful happening between physical hardware and cloud compute world. It's that now, hardware is revolutionizing to adopt AI-level computing speed. The acceleration in the last two years is a function of movement into cloud—AWS and Azure are getting better and Intel just created a new [6.2 Gigahertz processor](#). Hardware is getting specifically tailored to computing, and more acceleration is coming as more services matriculate downstream from huge company hypercomputing to large corporations and ultimately democratized in the entrepreneurial space. NVIDIA is now the 3rd or 4th most valuable company in the market cap. NVIDIA used to produce gaming processing units (GPU) which were rotated into multimodal, high-powered algorithmic compute. This is changing computing forever. NVIDIA was ridiculed for being in the gaming space and now all the GPU architecture is underpinning this large hyperscale cloud providers.

What are the new business risks being introduced as a result of higher computing speeds and AI?

The biggest business risk, which is also a societal risk, is that technology is now getting good enough to generate content that is nearly indistinguishable from the real thing. The sophistication of deep fakes will create a huge set of business risks related to authentication, identification, banking, personal information, etc.

The next associated risk is that then, the line between training data and product data will be erased—AI models will begin training on real world data which will change the way we manage new code and credentials. Industry will have to reimagine how they do business licenses and support authentication and identification.

Of course, the positive side is that companies can scale faster and produce more for less capital. But there will be risks that managing and regulatory bodies face as well.

What is a security solution that can help combat the rise of deep fakes.

At the moment, I don't know if there is anything concrete available, but lots of firms are aware of the risks. The only solution is to probably adopt AI on the mitigation side as well: to solve impossible problems or solve problems created by AI, the other side needs to adopt AI as well.

What are some effective regulations for AI usage?

At a high level, AI regulation will need to look at a wide range of issues from how we manage attack surface to vulnerability finder regulations to consumer data protection to the line between safety and privacy. Globally, AI and technology erases borders, so regulating AI will require governments and regulatory agencies to partner together to manage the AI-driven cyberspace.

Regulations such as tagging something as AI-generated may not be too useful: the process of tagging content is a function of humans looking at AI and trying to reduce it to a binary technology. Computers are binary—it uses 0's and 1's. But AI is modeled to be like a human and to not have a binary outcome. This means that attempting to tag AI-generated content will be far more complex because AI will learn to make exceptions and exclusions.

Regulation that focuses on training data may be more sophisticated and useful. If we understand the underlying data used in training, we can understand the core data and how AI chooses which data to make a decision. If we can tag and regulate the training data, we can manage risks by getting at the underlying risk factor and understanding how AI is trying to make predictions.

7. Appendix

Appendix 1: Table of Recent Examples of LLM-Themed TTPs by APTs²¹¹

APT	Description	TTPs
Forest Blizzard (also known as APT28, Fancy Bear)	<ul style="list-style-type: none"> Russian Military intelligence actor linked to GRU Unit 26165 Targeted victims of both tactical and strategic interest to the Russian government 	<ul style="list-style-type: none"> LLM-informed reconnaissance: used LLMs to research satellite and radar technologies specifically related to conventional military operations in Ukraine and broader research to bolster cyber operations. LLM-enhanced scripting techniques: used for basic scripting tasks such as file manipulation, data selection, regular expressions, and multiprocessing to potentially automate or optimize technical operations.
Emerald Sleet (THALLIUM)	<ul style="list-style-type: none"> North Korean threat actor Targeted prominent individuals with expertise on North Korea by impersonating reputable academic institutions and NGOs 	<ul style="list-style-type: none"> LLM-informed reconnaissance: used LLMs to identify think tanks, government organizations, or experts focusing on defense issues or North Korea’s nuclear weapons program. LLM-enhanced scripting techniques: employed LLMs for basic scripting tasks, such as programmatically identifying specific user events on a system. LLM-supported social engineering: used LLMs to assist in drafting and generating content likely used in spear-phishing campaigns targeting individuals with regional expertise. LLM-assisted vulnerability research: leveraged LLMs to understand publicly reported vulnerabilities, such as the CVE-2022-30190 Microsoft Support Diagnostic Tool (MSDT) vulnerability (known as “Follina”).
Crimson Sandstorm	<ul style="list-style-type: none"> Iranian threat actor allegedly connected to the Islamic Revolutionary Guard Corps (IRGC) Targeted various sectors including defense, maritime shipping, transportation, healthcare, and technology Relied on watering hole attacks and social engineering to deliver custom .NET malware 	<ul style="list-style-type: none"> LLM-enhanced scripting techniques: Used LLMs to generate code snippets supporting app and web development, remote server interactions, web scraping, task execution upon user sign-in, and sending system information via email. LLM-supported social engineering: Interacted with LLMs to create various phishing emails, including one posing as an international development agency and another enticing prominent feminists to an attacker-created website on feminism. LLM-enhanced anomaly detection evasion: Attempted to utilize LLMs to develop code for evading detection, learning methods to disable antivirus through registry or Windows policies, and deleting files in a directory after closing an application.
Charcoal Typhoon	<ul style="list-style-type: none"> Chinese state-affiliated threat actor Targeted government, higher education, communications infrastructure, oil & gas, and information technology Primarily targeted entities in Taiwan, Thailand, Mongolia, Malaysia, France, and Nepal, 	<ul style="list-style-type: none"> LLM-informed reconnaissance: Engaged LLMs to research and understand specific technologies, platforms, and vulnerabilities, indicating preliminary information-gathering stages. LLM-enhanced scripting techniques: Used LLMs to generate and refine scripts, aiming to streamline and automate complex cyber tasks and operations. LLM-supported social engineering: Leveraged LLMs for assistance with translations and communication, likely used to establish connections or manipulate targets. LLM-refined operational command techniques: Utilized LLMs for advanced commands, deeper system access, and

	specifically global institutions and individuals opposing China's policies	control reflecting post-compromise behavior.
Salmon Typhoon	<ul style="list-style-type: none"> • Chinese state-affiliated threat actor • Historically targeted US defense contractors, government agencies, and entities within the cryptographic technology sector 	<ul style="list-style-type: none"> • LLM-informed reconnaissance: Engaged LLMs for queries on various subjects, including global intelligence agencies, notable individuals, cybersecurity matters, and threat actors. • LLM-enhanced scripting techniques: Used LLMs to identify and fix coding errors, with requests for assistance in developing potentially malicious code, adhering to established ethical guidelines by declining such requests. • LLM-refined operational command techniques: Showed interest in specific file types and concealment tactics in operating systems, indicating efforts to refine operational command execution. • LLM-aided technical translation and explanation: Leveraged LLMs for translating computing terms and technical papers.

Appendix 2: Table of AI-Enabled Disinformation Efforts to Undermine Democracy and/or Increase Censorship²¹²

Country	Time period	Description
Pakistan	May 2023	<ul style="list-style-type: none"> Former Prime Minister Imran Khan shared an AI-generated video depicting a woman facing riot police to augment the narrative that Pakistani women supported him instead of the Pakistani military.
Nigeria	February 2023	<ul style="list-style-type: none"> AI-manipulated audio clip spread on social media purportedly implicating an opposition presidential candidate in plans to rig balloting.
United States	2023	<ul style="list-style-type: none"> Accounts affiliated with former President Donald Trump and Florida Governor Ron DeSantis shared videos with AI-generated content to undermine each other's candidacy. A manipulated video appeared on social media depicting President Biden making transphobic comments.
Venezuela	2023	<ul style="list-style-type: none"> State media outlets spread pro-government messages through AI-generated videos of news anchors from a fictional English-language channel produced by online AI tool, Synthesia. Graphika linked the company to a pro-CCP disinformation campaign targeting US audiences via the nonexistent news station "Wolf News."
India	2023	<ul style="list-style-type: none"> Prime Minister Modi and his Hindu nationalist Bharatiya Janata Party have incorporated censorship, including the use of AI-based automated systems, into the country's legal framework.
China	2023	<ul style="list-style-type: none"> Chatbots produced by China-based companies have refused to respond to user prompts on sensitive subjects such as Tiananmen Square and produce CCP claims about Taiwan. The Cyberspace Administration of China (CAC) has attempted to integrate CCP goals into the country's content recommendation algorithms, synthetic media, and generative AI tools. The CAC approved 41 suppliers of generative AI in mid-2023 and five chatbots were released to the public in August.
Russia	2022-ongoing	<ul style="list-style-type: none"> Private sector actors continue to spread disinformation about the Russian invasion of Ukraine through operations like "Doppelganger" and Cyber Front Z, which employ tactics such as mimicking Western media outlets and promoting anti-Ukraine propaganda.
Vietnam	2022	<ul style="list-style-type: none"> Authorities reportedly compelled Meta to remove all criticism of specified Communist Party of Vietnam (CPV) officials. The CPV also passed regulations to empower the Ministry of Public Security to prohibit platforms that do not comply with the requirement to remove toxic content within one day of notification.

8. Annotated Bibliography

The below annotated bibliography presents an overview of the ten key sources used in this report's literature review as well as their contribution to the report's analysis and key findings.

Aspen Digital. 'Envisioning Cyber Futures With AI'. Aspen Institute, 9 January 2024.
<https://www.aspendigital.org/report/cyber-futures-with-ai/>.

The report presents a timely analysis into potential future scenarios, challenges, and opportunities arising from the integration of AI into cybersecurity practices. The assessment draws on discussions with the Aspen Institute's US and Global Cybersecurity working groups to outline two extreme, yet realistic scenarios: a "good place" where AI tools disproportionately assist cybersecurity efforts and a "bad place" where attackers are instead advantaged by advancements in AI technology.

'Artificial Intelligence Index Report 2023'. Stanford University Human-Centered Artificial Intelligence, 2023. <https://aiindex.stanford.edu/report/>.

The annual 'Artificial Intelligence Index Report' is in its sixth year and aims to be the world's most credible and authoritative source for data and insights about AI. The report offers a comprehensive look into the latest trends and statistics in the field of AI, organized into eight main themes: research and development, technical performance, technical AI ethics, the economy, education, policy and governance, diversity, and public opinion. The index presents original data to provide a comprehensive review of the current state of AI advancement, and year-on-year comparisons.

Benaich, Nathan. 'State of AI Report'. Air Street Capital, 13 October 2023.
<https://www.stateof.ai/>.

The annual 'State of AI Report' is released by Air Street Capital, a venture capital firm investing in AI-first technology and life science companies. The report is now in its sixth year and is reviewed by leading AI practitioners in industry and research. It offers a compilation of data and analysis of AI developments along five key dimensions: research, industry, politics, safety, and predictions.

Funk, Allie, Adrian Shahbaz, and Kian Vesteinsson. 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence', 2023.

<https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence>.

This annual survey from Freedom House examines the state of internet freedom around the world. The 2023 edition considers the increasingly repressive applications of AI across the globe. The report provides detailed case studies where AI is being used to stifle dissent and suppress freedoms online. It also discusses the implications of these practices on internet freedom and democracy. The report is timely, and sheds lights on the evolving landscape of digital authoritarianism, enhanced by AI.

Hoffman, Wyatt, and Heeu Millie Kim. 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'. Center for Security and Emerging Technology, March 2023.

<https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/>.

This report addresses the nuanced challenges of integrating AI in military decision-making processes, providing a detailed context on applications in China and the US. The report presents recommendations aimed at mitigating these risks, ranging from developing AI-specific governance frameworks to promoting transparency and accountability in AI-assisted decisions. It serves as a critical resource for military strategists, policymakers, and researchers exploring the complexities of AI integration in defense strategies, emphasizing the need for responsible AI deployment for enhanced decision advantage while minimizing potential risks.

McKinsey. 'The State of AI in 2023: Generative AI's Breakout Year', 2023. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>.

This annual survey of global executives considers the current state of AI, highlighting the significant growth and potential of generative AI technologies and their increasing adoption across various sectors. The analysis outlines the transformative role generative AI plays in innovation, creativity, and problem-solving, and discusses the potential organizational risks from increased adoption. This report serves as a valuable resource for understanding the rapid evolution of generative AI and its implications for future technological landscapes.

Musser, Micah, Jonathan Spring, Christina Liaghati, Daniel Rohrer, Jonathan Elliott, Rumman Chowdhury, Andrew Lohn, et al. 'Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications'. Center for Security and Emerging Technology & Stanford Geopolitics, Technology and Governance Cyber Policy Center, April 2023. <https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/>.

The report addresses the distinct nature of AI vulnerabilities compared to traditional software vulnerabilities. It outlines recommendations endorsed by a diverse group of experts, ranging from incorporating AI vulnerabilities into cybersecurity frameworks to fostering collaboration between adversarial machine learning researchers and cybersecurity practitioners. This report serves as an important assessment of the complexities of securing AI systems against adversarial threats.

National Cyber Security Centre. 'The Near-Term Impact of AI on the Cyber Threat'. London, United Kingdom: National Cyber Security Centre, 24 January 2024. <https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>.

The NCSC Assessment (NCSC-A) report, released in January 2024, provides key insights on the impact of artificial intelligence (AI) on cybersecurity. The report provides predictions on how AI will amplify cyberattacks' frequency and impact over the next two years, especially in reconnaissance and social engineering. The report underscores the potential for AI to democratize cyber capabilities. To address these risks, the assessment advocates for continued investment in AI security, collaboration among cybersecurity communities, and vigilance against evolving AI-driven threats.

Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 'Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations'. NIST AI 100-2e2023. NIST Trustworthy and Responsible AI. Gaithersburg, MD: National Institute of Standards and Technology, January 2024. <https://doi.org/10.6028/NIST.AI.100-2e2023>.

This publication presents a taxonomy and terminology for understanding adversarial machine learning attacks and their mitigations. The paper was published as part of the NIST Trustworthy and Responsible AI series, providing a structured framework to categorize these attacks based on their objectives and techniques. The publication provides a clear and organized overview of the landscape of adversarial machine learning attacks and the corresponding defense mechanisms.

Sedova, Katerina, Christine McNeill, Aurora Johnson, and Aditi Joshi. 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model'. CSET Policy Brief. Center for Security and Emerging Technology (CSET), December 2021.

<https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/>.

This policy brief is the latest installment of a series that examines how advances in AI could be exploited to enhance operations that automate disinformation campaigns. The report describes how AI can supercharge current techniques to increase the speed, scale, and personalization of disinformation campaigns and examines how AI/ML technologies can enhance specific disinformation techniques and how these technologies may exacerbate current trends and shape future campaigns.

9. References

- 'AI Risk Management Framework'. *NIST*, 12 July 2021. <https://www.nist.gov/itl/ai-risk-management-framework>.
- 'AI Voice Generator & Text to Speech | ElevenLabs'. Accessed 20 February 2024. <https://elevenlabs.io/>.
- Antoniuk, Daryna. 'Russian Region Launches Chatbot to Report "Extremist" Neighbors', 30 November 2023. <https://therecord.media/russian-region-primorsky-krai-snitching-chatbot>.
- Appel, Gil, Juliana Neelbauer, and David A. Schweidel. 'Generative AI Has an Intellectual Property Problem'. *Harvard Business Review*, 7 April 2023. <https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem>.
- 'Artificial Intelligence Index Report 2023'. Stanford University Human-Centered Artificial Intelligence, 2023. <https://aiindex.stanford.edu/report/>.
- Aspen Digital. 'Envisioning Cyber Futures With AI'. Aspen Institute, 9 January 2024. <https://www.aspendigital.org/report/cyber-futures-with-ai/>.
- . 'Generative A.I. Regulation and Cybersecurity: A Global View of Policymaking'. Aspen Institute, 16 January 2024. <https://www.aspendigital.org/report/generative-ai-regulation-and-cybersecurity/>.
- Batalis, Steph. 'AI and Biorisk: An Explainer'. Center for Security and Emerging Technology, December 2023. <https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer/>.
- Benaich, Nathan. 'State of AI Report'. Air Street Capital, 13 October 2023. <https://www.stateof.ai/>.
- Bergengruen, Vera. 'How Tech Giants Turned Ukraine Into an AI War Lab'. *Time*, 8 February 2024. <https://time.com/magazine/us/6695603/february-26th-2024-vol-203-no-5-worldwide/>.
- Berman, Noah, Lindsay Maizland, and Andrew Chatzky. 'Is China's Huawei a Threat to U.S. National Security?' Council on Foreign Relations, 8 February 2023. <https://www.cfr.org/backgrounder/chinas-huawei-threat-us-national-security>.
- Bora, Sanjeev. 'The Rise of Generative AI in Audio, Revolutionizing Music Production: "Stable Audio" Kicking the Competition'. *Medium* (blog), 17 September 2023. <https://medium.com/@sanjeeva.bora/the-rise-of-generative-ai-in-audio-revolutionizing-music-production-stable-audio-kicking-the-8ee289f1616f>.
- Borgonovo, Federico, Silvano Rizieri Lucini, and Giulia Porrino. 'Weapons of Mass Hate Dissemination: The Use of Artificial Intelligence by Right-Wing Extremists'. *GNET* (blog), 23 February 2024. <https://gnet-research.org/2024/02/23/weapons-of-mass-hate-dissemination-the-use-of-artificial-intelligence-by-right-wing-extremists/>.
- CIO Dive. 'AI-Generated Code Leads to Security Issues for Most Businesses: Report'. Accessed 7 March 2024. <https://www.ciodive.com/news/security-issues-ai-generated-code-snyk/705900/>.
- Code42. 'Annual Data Exposure Report 2024', 2024. <https://www.code42.com/resources/promo-resources/2024-data-exposure>.
- Cybersecurity & Infrastructure Security Agency (CISA). 'Foreign Influence Operations and Disinformation | Cybersecurity and Infrastructure Security Agency CISA', 2023. <https://www.cisa.gov/topics/election-security/foreign-influence-operations-and-disinformation>.
- Cybersecurity and Infrastructure Security Agency (CISA). 'Risk in Focus: Generative A.I. and the 2024 Election Cycle'. Cybersecurity & Infrastructure Security Agency (CISA), 18 January 2024.
- 'Data Poisoning Attacks: A New Attack Vector within AI | Cobalt'. Accessed 26 February 2024. <https://www.cobalt.io/blog/data-poisoning-attacks-a-new-attack-vector-within-ai>.
- Deng, Gelei, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 'MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots'. In *Proceedings 2024 Network and Distributed System Security Symposium*, 2024. <https://doi.org/10.14722/ndss.2024.24188>.

Dutta, Tushar Subhra. 'Hackers Released New Black Hat AI Tools XXXGPT and Wolf GPT'. Cyber Security News, 1 August 2023. <https://cybersecuritynews.com/black-hat-ai-tools-xxxgpt-and-wolf-gpt/>.

Erzberger, Arthur. 'WormGPT and FraudGPT – The Rise of Malicious LLMs', 8 August 2023. <https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/>.

Fedasiuk, Ryan, Jennifer Melot, and Ben Murphy. 'Harnessed Lightning'. Center for Security and Emerging Technology, October 2021. <https://cset.georgetown.edu/publication/harnessed-lightning/>.

Feldstein, Steven. 'AI in War: Can Advanced Military Technologies Be Tamed before It's Too Late?' *Bulletin of the Atomic Scientists* (blog), 11 January 2024. <https://thebulletin.org/2024/01/ai-in-war-can-advanced-military-technologies-be-tamed-before-its-too-late/>.

Funk, Allie, Adrian Shahbaz, and Kian Vesteinsson. 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence', 2023. <https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence>.

Gentile, Gian, Michael Shurkin, Alexandra T. Evans, Michelle Gris , Mark Hvizda, and Rebecca Jensen. 'A History of the Third Offset, 2014–2018'. RAND Corporation, 31 March 2021. https://www.rand.org/pubs/research_reports/RRA454-1.html.

Grinnell, Rick. 'CIOs Are Worried about the Informal Rise of Generative AI in the Enterprise'. CIO, 30 August 2023. <https://www.cio.com/article/650764/cios-are-worried-about-the-informal-rise-of-generative-ai-in-the-enterprise.html>.

Hassan. 'How AI Will Change the Future of Mass Surveillance'. Cybernews, 31 December 2023. <https://cybernews.com/editorial/how-ai-will-change-the-future-of-mass-surveillance/>.

'HEARING TO RECEIVE TESTIMONY ON ARTIFICIAL INTELLIGENCE APPLICATIONS TO OPERATIONS IN CYBERSPACE'. Washington D.C., 5 March 2022.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. 'An Overview of Catastrophic AI Risks', 2023. <https://arxiv.org/abs/2306.12001>.

Hoffman, Wyatt, and Heeu Millie Kim. 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'. Center for Security and Emerging Technology, March 2023. <https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/>.

Honigberg, Bradley. 'The Existential Threat of AI-Enhanced Disinformation Operations'. Just Security, 8 July 2022. <https://www.justsecurity.org/82246/the-existential-threat-of-ai-enhanced-disinformation-operations/>.

Hsu, Tiffany. 'Fake and Explicit Images of Taylor Swift Started on 4chan, Study Says'. *The New York Times*, 5 February 2024, sec. Business. <https://www.nytimes.com/2024/02/05/business/media/taylor-swift-ai-fake-images.html>.

IBM. 'What Is Semi-Supervised Learning?' Accessed 1 March 2024. <https://www.ibm.com/topics/semi-supervised-learning>.

IBM Data and AI Team. 'Shedding Light on AI Bias with Real World Examples'. IBM Blog, 16 October 2023. <https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples>.

Insikt Group. 'Aggressive Malign Influence Threatens to Shape US 2024 Elections'. Recorded Future, 14 December 2023. <https://www.recordedfuture.com/aggressive-malign-influence-threatens-us-2024-elections>.

Jung, Col. 'AI Revolution — Your Fast-Paced Introduction to Machine Learning'. Medium, 22 April 2023. <https://medium.com/geekculture/ai-revolution-your-fast-paced-introduction-to-machine-learning-914ce9b6ddf>.

Keast, Jacinta. 'Shadow Play: A pro-China and Anti-US Influence Operation Thrives on YouTube'. The Strategist, 14 December 2023. <https://www.aspistrategist.org.au/shadow-play-a-pro-china-and-anti-us-influence-operation-thrives-on-youtube/>.

- Kelley, Daniel. 'WormGPT - The Generative AI Tool Cybercriminals Are Using to Launch BEC Attacks'. *SlashNext* (blog), 13 July 2023. <https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/>.
- Klepper, David, and Huizhong Wu. 'How Taiwan Beat Back Disinformation and Preserved the Integrity of Its Election'. AP News, 27 January 2024. <https://apnews.com/article/taiwan-election-china-disinformation-vote-fraud-4968ef08fd13821e359b8e195b12919c>.
- Knibbs, Kate. 'Researchers Say the Deepfake Biden Robocall Was Likely Made With Tools From AI Startup ElevenLabs'. *Wired*. Accessed 20 February 2024. <https://www.wired.com/story/biden-robocall-deepfake-elevenlabs/>.
- Lakatos, Santiago. 'A Revealing Picture'. Graphika, December 2023. <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf>.
- Lakomy, Miron. 'Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities'. *GNET* (blog), 15 December 2023. <https://gnet-research.org/2023/12/15/artificial-intelligence-as-a-terrorism-enabler-understanding-the-potential-impact-of-chatbots-and-image-generators-on-online-terrorist-activities/>.
- Lark Editorial Team. 'Benchmarking', 27 December 2023. https://www.larksuite.com/en_us/topics/ai-glossary/benchmarking.
- Lee, Nicol Turner, and Caitlin Chin-Rothmann. 'Police Surveillance and Facial Recognition: Why Data Privacy Is Imperative for Communities of Color'. *Brookings* (blog), 12 April 2022. <https://www.brookings.edu/articles/police-surveillance-and-facial-recognition-why-data-privacy-is-an-imperative-for-communities-of-color/>.
- Mascellino, Alessandro. 'Dark Web Markets Offer New FraudGPT AI Tool'. *Infosecurity Magazine*, 26 July 2023. <https://www.infosecurity-magazine.com/news/dark-web-markets-fraudgpt-ai-tool/>.
- McKinsey. 'The Economic Potential of Generative AI: The next Productivity Frontier', 14 June 2023. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction>.
- . 'The State of AI in 2023: Generative AI's Breakout Year'. , 2023. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>.
- McMillan, Robert, Dustin Volz, and Aruna Viswanatha. 'China Is Stealing AI Secrets to Turbocharge Spying, U.S. Says - WSJ'. *The Wall Street Journal*, 25 December 2023. <https://www.wsj.com/tech/ai/china-is-stealing-ai-secrets-to-turbocharge-spying-u-s-says-00413594>.
- Menlo Security. 'The Continued Impact of Generative AI on Security Posture', 2024.
- Microsoft Corporation and Berkman Klein Center for Internet and Society at Harvard University. 'Failure Modes in Machine Learning', 2 November 2022. <https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning>.
- Microsoft Threat Intelligence. 'Staying Ahead of Threat Actors in the Age of AI'. Microsoft Security Blog, 14 February 2024. <https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/>.
- Montalbano, Elizabeth. "'DarkBERT" GPT-Based Malware Trains Up on the Entire Dark Web', 1 August 2023. <https://www.darkreading.com/application-security/gpt-based-malware-trains-dark-web>.
- Mouton, Christopher A., Caleb Lucas, and Ella Guest. 'The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach'. RAND Corporation, 16 October 2023. https://www.rand.org/pubs/research_reports/RRA2977-1.html.
- Mukherjee, Supantha. 'Spotify to Use Google's AI to Tailor Podcasts, Audiobooks Recommendations'. *Reuters*, 16 November 2023, sec. Technology.

- <https://www.reuters.com/technology/spotify-use-googles-ai-tailor-podcasts-audiobooks-recommendations-2023-11-16/>.
- Musser, Micah, Jonathan Spring, Christina Liaghati, Daniel Rohrer, Jonathan Elliott, Rumman Chowdhury, Andrew Lohn, et al. 'Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications'. Center for Security and Emerging Technology & Stanford Geopolitics, Technology and Governance Cyber Policy Center, April 2023.
<https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/>.
- Mysla, Vlad. 'Machine Learning in 10 Minutes'. Medium, 22 April 2020.
<https://medium.datadriveninvestor.com/machine-learning-in-10-minutes-354d83e5922e>.
- National Cyber Security Centre. 'The Near-Term Impact of AI on the Cyber Threat'. London, United Kingdom: National Cyber Security Centre, 24 January 2024.
<https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>.
- NIST. 'U.S. Artificial Intelligence Safety Institute', 26 October 2023. <https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute>.
- Office of the Privacy Commissioner of Canada. 'PIPEDA Findings #2021-001: Joint Investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information Du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta', 3 February 2021.
<https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/>.
- Ozair, Merav and PhD. 'Misinformation in the Age of Artificial Intelligence and What It Means for the Markets'. Nasdaq, 22 November 2023. <https://www.nasdaq.com/articles/misinformation-in-the-age-of-artificial-intelligence-and-what-it-means-for-the-markets>.
- Perry, Neil, Megha Srivastava, Deepak Kumar, and Dan Boneh. 'Do Users Write More Insecure Code with AI Assistants?' In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2785–99. CCS '23. New York, NY, USA: Association for Computing Machinery, 2023. <https://doi.org/10.1145/3576915.3623157>.
- Poireault, Kevin. 'The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web', 10 August 2023. <https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html>.
- Ramponi. 'Recent Developments in Generative AI for Audio'. *AssemblyAI* (blog), 27 June 2023.
<https://www.assemblyai.com/blog/recent-developments-in-generative-ai-for-audio/>.
- Ryan-Mosley, Tate. 'How Generative AI Is Boosting the Spread of Disinformation and Propaganda | MIT Technology Review'. MIT Technology Review, 4 October 2023.
<https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house>.
- Scharre, Paul. 'The Perilous Coming Age of AI Warfare'. *Foreign Affairs*, 29 February 2024.
<https://www.foreignaffairs.com/ukraine/perilous-coming-age-ai-warfare>.
- Scharre, Paul, and Aaron Mehta. 'How AI Became "the Autocrat's New Toolkit" [BOOK EXCERPT]'. *Breaking Defense* (blog), 28 February 2023.
<https://breakingdefense.sites.breakingmedia.com/2023/02/how-ai-became-the-autocrats-new-toolkit-book-excerpt/>.
- Sedova, Katerina, Christine McNeill, Aurora Johnson, and Aditi Joshi. 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model'. CSET Policy Brief. Center for Security and Emerging Technology (CSET), December 2021.
<https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/>.
- Sharma, Shweta. 'ChatGPT Creates Mutating Malware That Evades Detection by EDR'. CSO Online, 6 June 2023. <https://www.csoonline.com/article/575487/chatgpt-creates-mutating-malware-that-evades-detection-by-edr.html>.

- . ‘Samsung Bans Staff AI Use over Data Leak Concerns’. CSO Online, 2 May 2023. <https://www.csoonline.com/article/575215/samsung-bans-staff-ai-use-over-data-leak-concerns.html>.
- ‘Sora’. Accessed 17 February 2024. <https://openai.com/sora>.
- Tech Against Terrorism. ‘Early Terrorist Experimentation with Generative Artificial Intelligence Services’, November 2023. <https://techagainstterrorism.org/news/early-terrorist-adoption-of-generative-ai>.
- ‘The State of Phishing 2023’. SlashNext, 2023.
- The White House. ‘Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence’. The White House, 30 October 2023. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
- Thompson, Stuart A., and Sapna Maheshwari. ‘“A.I. Obama” and Fake Newscasters: How A.I. Audio Is Swarming TikTok’. *The New York Times*, 12 October 2023, sec. Technology. <https://www.nytimes.com/2023/10/12/technology/tiktok-ai-generated-voices-disinformation.html>.
- Toner, Helen. ‘What Are Generative AI, Large Language Models, and Foundation Models?’ *Center for Security and Emerging Technology* (blog), 12 May 2023. <https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/>.
- United States Department of Justice. ‘Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google’, 6 March 2024. <https://www.justice.gov/opa/pr/chinese-national-residing-california-arrested-theft-artificial-intelligence-related-trade>.
- Vassilev, Apostol, Alina Oprea, Alie Fordyce, and Hyrum Anderson. ‘Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations’. NIST AI 100-2e2023. NIST Trustworthy and Responsible AI. Gaithersburg, MD: National Institute of Standards and Technology, January 2024. <https://doi.org/10.6028/NIST.AI.100-2e2023>.
- Vincent, Brandi. ‘Scale AI to Set the Pentagon’s Path for Testing and Evaluating Large Language Models’. *DefenseScoop* (blog), 20 February 2024. <https://defensescoop.com/2024/02/20/scale-ai-pentagon-testing-evaluating-large-language-models/>.
- VMWare. ‘Global Incident Response: Threat Report’. Palo Alto, CA, 2022. https://www.vmware.com/content/dam/learn/en/amer/fy23/pdf/1553238_Global_Incident_Response_Threat_Report_Weathering_The_Storm.pdf.
- Woollacott, Emma. ‘China Targets US Voters With New AI Misinformation Techniques’. *Forbes*, 8 September 2023, sec. Cybersecurity. <https://www.forbes.com/sites/emmawoollacott/2023/09/08/china-targets-us-voters-with-new-ai-misinformation-techniques/>.
- World Economic Forum. ‘Global Risks Report 2024’, 10 January 2024. <https://www.weforum.org/publications/global-risks-report-2024/in-full/>.

10. Footnotes

- ¹ McKinsey, 'The Economic Potential of Generative AI: The next Productivity Frontier'.
- ² NIST, 'U.S. Artificial Intelligence Safety Institute'.
- ³ Drawn from various sources including: Jung, 'AI Revolution — Your Fast-Paced Introduction to Machine Learning'; Mysla, 'Machine Learning in 10 Minutes'.
- ⁴ A chief architect Interview.
- ⁵ Benaich, 'State of AI Report'; IBM, 'What Is Semi-Supervised Learning?'; 'Artificial Intelligence Index Report 2023'.
- ⁶ Toner, 'What Are Generative AI, Large Language Models, and Foundation Models?'
- ⁷ McKinsey, 'The State of AI in 2023: Generative AI's Breakout Year'.
- ⁸ In ML models, parameters are numerical values that are learned during training. The value of parameters in machine learning models determines how a model might interpret input data and make predictions.
- ⁹ 'Artificial Intelligence Index Report 2023', 11.
- ¹⁰ 'Artificial Intelligence Index Report 2023', 11.
- ¹¹ 'Artificial Intelligence Index Report 2023', 11.
- ¹² 'Artificial Intelligence Index Report 2023', 50.
- ¹³ Significant machine learning systems refers to large language and multimodal models.
- ¹⁴ 'Artificial Intelligence Index Report 2023', 51.
- ¹⁵ 'Artificial Intelligence Index Report 2023'.
- ¹⁶ A chief architect Interview.
- ¹⁷ McKinsey, 'The State of AI in 2023: Generative AI's Breakout Year'.
- ¹⁸ McKinsey.
- ¹⁹ McKinsey.
- ²⁰ McKinsey. Source: 2023 McKinsey Global survey
- ²¹ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ²² Ramponi, 'Recent Developments in Generative AI for Audio'.
- ²³ Bora, 'The Rise of Generative AI in Audio, Revolutionizing Music Production'.
- ²⁴ Mukherjee, 'Spotify to Use Google's AI to Tailor Podcasts, Audiobooks Recommendations'.
- ²⁵ 'AI Voice Generator & Text to Speech | ElevenLabs'.
- ²⁶ Knibbs, 'Researchers Say the Deepfake Biden Robocall Was Likely Made With Tools From AI Startup ElevenLabs'.
- ²⁷ Thompson and Maheshwari, "'A.I. Obama" and Fake Newscasters'.
- ²⁸ Thompson and Maheshwari, "'A.I. Obama" and Fake Newscasters'.
- ²⁹ 'Artificial Intelligence Index Report 2023', 98.
- ³⁰ 'Sora'.
- ³¹ VMWare, 'Global Incident Response: Threat Report'.
- ³² Hsu, 'Fake and Explicit Images of Taylor Swift Started on 4chan, Study Says'.
- ³³ Appel, Neelbauer, and Schweidel, 'Generative AI Has an Intellectual Property Problem'.
- ³⁴ Keast, 'Shadow Play'.
- ³⁵ Insikt Group, 'Aggressive Malign Influence Threatens to Shape US 2024 Elections'.
- ³⁶ 'The State of Phishing 2023'.
- ³⁷ 'The State of Phishing 2023', 8.
- ³⁸ Kelley, 'WormGPT - The Generative AI Tool Cybercriminals Are Using to Launch BEC Attacks'.
- ³⁹ Erzberger, 'WormGPT and FraudGPT – The Rise of Malicious LLMs'.
- ⁴⁰ Mascellino, 'Dark Web Markets Offer New FraudGPT AI Tool'.
- ⁴¹ Poireault, 'The Dark Side of Generative AI'.
- ⁴² Dutta, 'Hackers Released New Black Hat AI Tools XXXGPT and Wolf GPT'.
- ⁴³ Montalbano, "'DarkBERT" GPT-Based Malware Trains Up on the Entire Dark Web'.
- ⁴⁴ 'The State of Phishing 2023', 12.
- ⁴⁵ Deng et al., 'MasterKey'.
- ⁴⁶ Deng et al.
- ⁴⁷ Anonymous Expert 2 Interview.
- ⁴⁸ National Cyber Security Centre, 'The Near-Term Impact of AI on the Cyber Threat'.
- ⁴⁹ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁵⁰ Aspen Digital.

-
- ⁵¹ A cyber security company CEO Interview.
- ⁵² A cyber security company CEO Interview.
- ⁵³ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁵⁴ Microsoft Threat Intelligence, 'Staying Ahead of Threat Actors in the Age of AI'.
- ⁵⁵ Microsoft Threat Intelligence.
- ⁵⁶ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁵⁷ Sharma, 'ChatGPT Creates Mutating Malware That Evades Detection by EDR'.
- ⁵⁸ Adapted from: Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁵⁹ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁶⁰ Aspen Digital.
- ⁶¹ Aspen Digital.
- ⁶² Aspen Digital, 10.
- ⁶³ 'The State of Phishing 2023', 1.
- ⁶⁴ 'The State of Phishing 2023', 1.
- ⁶⁵ 'The State of Phishing 2023', 1.
- ⁶⁶ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁶⁷ National Cyber Security Centre, 'The Near-Term Impact of AI on the Cyber Threat'.
- ⁶⁸ National Cyber Security Centre.
- ⁶⁹ Aspen Digital, 'Envisioning Cyber Futures With AI'.
- ⁷⁰ National Cyber Security Centre, 'The Near-Term Impact of AI on the Cyber Threat'.
- ⁷¹ World Economic Forum, 'Global Risks Report 2024', 18.
- ⁷² Sedova et al., 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model'.
- ⁷³ Cybersecurity & Infrastructure Security Agency (CISA), 'Foreign Influence Operations and Disinformation | Cybersecurity and Infrastructure Security Agency CISA'.
- ⁷⁴ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ⁷⁵ Honigberg, 'The Existential Threat of AI-Enhanced Disinformation Operations'.
- ⁷⁶ Honigberg.
- ⁷⁷ Sedova et al., 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model', 7.
- ⁷⁸ Honigberg, 'The Existential Threat of AI-Enhanced Disinformation Operations'.
- ⁷⁹ Honigberg.
- ⁸⁰ Sedova et al., 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model', 45.
- ⁸¹ Cybersecurity and Infrastructure Security Agency (CISA), 'Risk in Focus: Generative A.I. and the 2024 Election Cycle'.
- ⁸² Ryan-Mosley, 'How Generative AI Is Boosting the Spread of Disinformation and Propaganda | MIT Technology Review'.
- ⁸³ Honigberg, 'The Existential Threat of AI-Enhanced Disinformation Operations'.
- ⁸⁴ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ⁸⁵ World Economic Forum, 'Global Risks Report 2024', 20.
- ⁸⁶ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ⁸⁷ Funk, Shahbaz, and Vesteinsson.
- ⁸⁸ Funk, Shahbaz, and Vesteinsson.
- ⁸⁹ Funk, Shahbaz, and Vesteinsson.
- ⁹⁰ Antoniuk, 'Russian Region Launches Chatbot to Report "Extremist" Neighbors'.
- ⁹¹ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ⁹² Sedova et al., 'AI and the Future of Disinformation Campaigns: Part 2: A Threat Model'.
- ⁹³ World Economic Forum, 'Global Risks Report 2024'.
- ⁹⁴ Lakatos, 'A Revealing Picture'.
- ⁹⁵ Lakatos, 1.
- ⁹⁶ Lakatos, 'A Revealing Picture'.
- ⁹⁷ Ozair and PhD, 'Misinformation in the Age of Artificial Intelligence and What It Means for the Markets'.
- ⁹⁸ Ozair and PhD.
- ⁹⁹ World Economic Forum, 'Global Risks Report 2024', 21.
- ¹⁰⁰ World Economic Forum, 21.
- ¹⁰¹ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ¹⁰² World Economic Forum, 'Global Risks Report 2024', 19.
- ¹⁰³ Klepper and Wu, 'How Taiwan Beat Back Disinformation and Preserved the Integrity of Its Election'.
- ¹⁰⁴ Klepper and Wu.

-
- ¹⁰⁵ World Economic Forum, 'Global Risks Report 2024', 19; Woollacott, 'China Targets US Voters With New AI Misinformation Techniques'.
- ¹⁰⁶ Cybersecurity and Infrastructure Security Agency (CISA), 'Risk in Focus: Generative A.I. and the 2024 Election Cycle'.
- ¹⁰⁷ Musser et al., 'Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications', 6; 'HEARING TO RECEIVE TESTIMONY ON ARTIFICIAL INTELLIGENCE APPLICATIONS TO OPERATIONS IN CYBERSPACE'.
- ¹⁰⁸ Vassilev et al., 'Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations'.
- ¹⁰⁹ 'Data Poisoning Attacks'.
- ¹¹⁰ 'Data Poisoning Attacks'; Vassilev et al., 'Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations'.
- ¹¹¹ Taken from: Microsoft Corporation and Berkman Klein Center for Internet and Society at Harvard University, 'Failure Modes in Machine Learning'.
- ¹¹² Taken from: Microsoft Corporation and Berkman Klein Center for Internet and Society at Harvard University.
- ¹¹³ Microsoft Corporation and Berkman Klein Center for Internet and Society at Harvard University.
- ¹¹⁴ Taken from: Microsoft Corporation and Berkman Klein Center for Internet and Society at Harvard University.
- ¹¹⁵ Musser et al., 'Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications', 8.
- ¹¹⁶ IBM Data and AI Team, 'Shedding Light on AI Bias with Real World Examples'.
- ¹¹⁷ IBM Data and AI Team.
- ¹¹⁸ IBM Data and AI Team.
- ¹¹⁹ Vassilev et al., 'Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations', 3.
- ¹²⁰ Vassilev et al., 3.
- ¹²¹ 'Data Poisoning Attacks'.
- ¹²² Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage', 7–8.
- ¹²³ Vincent, 'Scale AI to Set the Pentagon's Path for Testing and Evaluating Large Language Models'.
- ¹²⁴ Vincent.
- ¹²⁵ Benaich, 'State of AI Report'.
- ¹²⁶ Bergengruen, 'How Tech Giants Turned Ukraine Into an AI War Lab'.
- ¹²⁷ Bergengruen.
- ¹²⁸ Fully autonomous weapon, or lethal autonomous weapons (LAWS) are a specific type of weapons systems that use sensors and ML algorithms to identify targets and independently deploy an onboard weapon system to engage and destroy the target without manual human control.
- ¹²⁹ Scharre, 'The Perilous Coming Age of AI Warfare'.
- ¹³⁰ Feldstein, 'AI in War'.
- ¹³¹ Feldstein.
- ¹³² Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'.
- ¹³³ Hoffman and Kim, 12.
- ¹³⁴ Fedasuk, Melot, and Murphy, 'Harnessed Lightning'.
- ¹³⁵ Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage', 12.
- ¹³⁶ Gentile et al., 'A History of the Third Offset, 2014–2018'.
- ¹³⁷ Benaich, 'State of AI Report'.
- ¹³⁸ Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage', 12–15.
- ¹³⁹ Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'.
- ¹⁴⁰ Hoffman and Kim.
- ¹⁴¹ Hoffman and Kim.
- ¹⁴² McMillan, Volz, and Viswanatha, 'China Is Stealing AI Secrets to Turbocharge Spying, U.S. Says - WSJ'.
- ¹⁴³ McMillan, Volz, and Viswanatha.
- ¹⁴⁴ 'HEARING TO RECEIVE TESTIMONY ON ARTIFICIAL INTELLIGENCE APPLICATIONS TO OPERATIONS IN CYBERSPACE'.
- ¹⁴⁵ Hassan, 'How AI Will Change the Future of Mass Surveillance'.
- ¹⁴⁶ Hassan.
- ¹⁴⁷ Hassan.
- ¹⁴⁸ Hassan.
- ¹⁴⁹ Hassan.
- ¹⁵⁰ Hassan.
- ¹⁵¹ Tech Against Terrorism, 'Early Terrorist Experimentation with Generative Artificial Intelligence Services'.
- ¹⁵² Tech Against Terrorism.

-
- ¹⁵³ Borgonovo, Rizieri Lucini, and Porrino, 'Weapons of Mass Hate Dissemination'.
- ¹⁵⁴ Borgonovo, Rizieri Lucini, and Porrino.
- ¹⁵⁵ Borgonovo, Rizieri Lucini, and Porrino.
- ¹⁵⁶ The White House, 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'.
- ¹⁵⁷ Mouton, Lucas, and Guest, 'The Operational Risks of AI in Large-Scale Biological Attacks'.
- ¹⁵⁸ Hendrycks, Mazeika, and Woodside, 'An Overview of Catastrophic AI Risks'.
- ¹⁵⁹ Hendrycks, Mazeika, and Woodside.
- ¹⁶⁰ Hendrycks, Mazeika, and Woodside.
- ¹⁶¹ Scharre and Mehta, 'How AI Became "the Autocrat's New Toolkit" [BOOK EXCERPT]'.
- ¹⁶² Scharre and Mehta, 'How AI Became "the Autocrat's New Toolkit" [BOOK EXCERPT]'.
- ¹⁶³ Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.
- ¹⁶⁴ Scharre and Mehta, 'How AI Became "the Autocrat's New Toolkit" [BOOK EXCERPT]'.
- ¹⁶⁵ Berman, Maizland, and Chatzky, 'Is China's Huawei a Threat to U.S. National Security?'
- ¹⁶⁶ Lee and Chin-Rothmann, 'Police Surveillance and Facial Recognition'.
- ¹⁶⁷ Lee and Chin-Rothmann.
- ¹⁶⁸ Office of the Privacy Commissioner of Canada, 'PIPEDA Findings #2021-001'.
- ¹⁶⁹ Lee and Chin-Rothmann, 'Police Surveillance and Facial Recognition'.
- ¹⁷⁰ Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage', 1.
- ¹⁷¹ Hoffman and Kim, 1.
- ¹⁷² Hoffman and Kim, 1.
- ¹⁷³ Hoffman and Kim, 1.
- ¹⁷⁴ Bergengruen, 'How Tech Giants Turned Ukraine Into an AI War Lab'.
- ¹⁷⁵ Bergengruen.
- ¹⁷⁶ Bergengruen.
- ¹⁷⁷ Bergengruen.
- ¹⁷⁸ Feldstein, 'AI in War'.
- ¹⁷⁹ Lakomy, 'Artificial Intelligence as a Terrorism Enabler?'
- ¹⁸⁰ Lakomy.
- ¹⁸¹ Lakomy.
- ¹⁸² Lakomy.
- ¹⁸³ Mouton, Lucas, and Guest, 'The Operational Risks of AI in Large-Scale Biological Attacks'; Batalis, 'AI and Biorisk: An Explainer'.
- ¹⁸⁴ 'Artificial Intelligence Index Report 2023', 14.
- ¹⁸⁵ McKinsey, 'The State of AI in 2023: Generative AI's Breakout Year'.
- ¹⁸⁶ McKinsey.
- ¹⁸⁷ McKinsey.
- ¹⁸⁸ McKinsey.
- ¹⁸⁹ Perry et al., 'Do Users Write More Insecure Code with AI Assistants?'; 'AI-Generated Code Leads to Security Issues for Most Businesses'.
- ¹⁹⁰ Perry et al., 'Do Users Write More Insecure Code with AI Assistants?'
- ¹⁹¹ Grinnell, 'CIOs Are Worried about the Informal Rise of Generative AI in the Enterprise'.
- ¹⁹² Grinnell.
- ¹⁹³ Sharma, 'Samsung Bans Staff AI Use over Data Leak Concerns'.
- ¹⁹⁴ Sharma.
- ¹⁹⁵ McKinsey, 'The State of AI in 2023: Generative AI's Breakout Year'.
- ¹⁹⁶ Appel, Neelbauer, and Schweidel, 'Generative AI Has an Intellectual Property Problem'.
- ¹⁹⁷ Appel, Neelbauer, and Schweidel, 'Generative AI Has an Intellectual Property Problem'.
- ¹⁹⁸ Aspen Digital, 'Generative A.I. Regulation and Cybersecurity: A Global View of Policymaking'; 'AI Risk Management Framework'.
- ¹⁹⁹ Menlo Security, 'The Continued Impact of Generative AI on Security Posture'.
- ²⁰⁰ Menlo Security.
- ²⁰¹ Code42, 'Annual Data Exposure Report 2024'.
- ²⁰² United States Department of Justice, 'Chinese National Residing in California Arrested for Theft of Artificial Intelligence-Related Trade Secrets from Google'.
- ²⁰³ United States Department of Justice.
- ²⁰⁴ A national security expert Interview.

²⁰⁵ A national security expert Interview.

²⁰⁶ 'Artificial Intelligence Index Report 2023'.

²⁰⁷ Vassilev et al., 'Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations', 4.

²⁰⁸ Menlo Security, 'The Continued Impact of Generative AI on Security Posture'.

²⁰⁹ National Cyber Security Centre, 'The Near-Term Impact of AI on the Cyber Threat'.

²¹⁰ Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'.

²¹¹ Microsoft Threat Intelligence, 'Staying Ahead of Threat Actors in the Age of AI'.

²¹² Funk, Shahbaz, and Vesteinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence'.