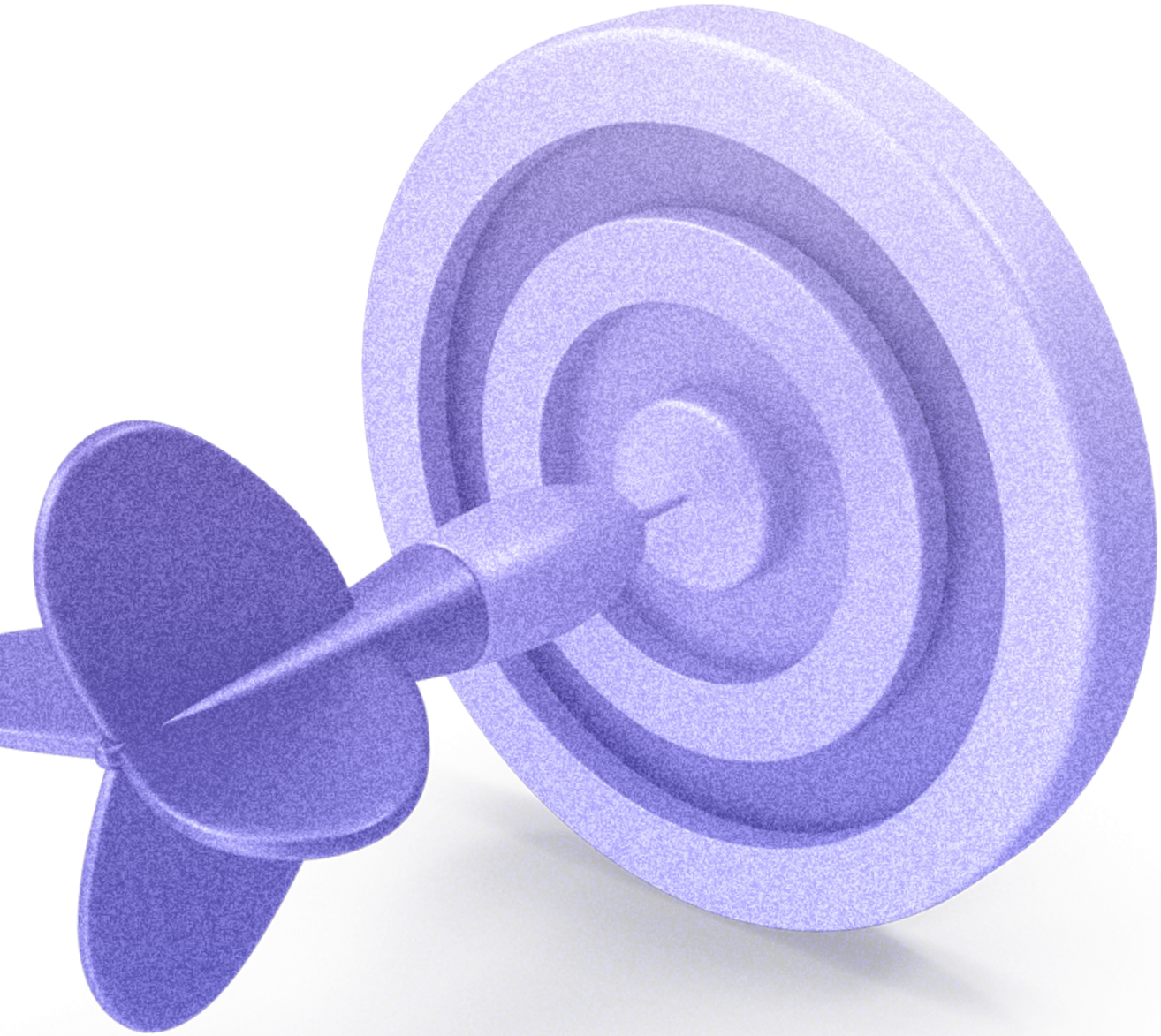
FS-ISAC

# Combating Threats and Reducing Risks Posed by AI



February 2024

# Contents

## EXECUTIVE SUMMARY

The mission of this working group is to provide recommendations and best practices to the financial services industry to combat cybersecurity threats and reduce the risks posed by artificial intelligence (AI).

While there are many beneficial uses of AI in financial services, threat actors are using AI in their campaigns and cyber attacks. And though most models contain safeguards to prevent malicious use, researchers have demonstrated time and time again how cunning and creativity can circumvent these controls to achieve malign intent.

This paper highlights a few of the areas where threat actors, from nation-state actors to cyber criminals, are endeavoring to exploit AI for nefarious purposes, including:

- The offensive use of AI to target people, processes, and technology.
- The adversarial targeting of AI-based solutions.
- Risks within the AI software supply chain.
- Inadvertent risks arising from the use of generative AI.
- External factors, including legal, regulatory, and ethical threats.

Artificial intelligence is constantly evolving and so are attack campaigns. This work, therefore, is not meant to be comprehensive. Rather, it is designed to arm cybersecurity experts in financial services with information and defense techniques pertinent to the threat landscape they face now.

## Social Engineering and Phishing Emails

Currently, the assessment of the security community is that while adversaries are experimenting with leveraging AI, most operational usage remains limited and primarily related to social engineering.

The security community has long assessed that AI can and will be leveraged to craft more realistic and convincing phishing lures for social engineering, spear phishing campaigns, business email compromise (BEC) messages, and other scams.

This is possible despite multiple safeguards. As FBI Director Christopher Wray recently explained, "What if I tell [an AI model] to write a formal business email, from one banking employee to another, to instruct them to wire money and ensure the coworker understands that the request is urgent?" The AI output is one that, with a few simple tweaks, can be used by fraudsters to conduct legitimate BEC compromises. Wray further explained that paired with AI-generated images, fraud campaigns will become even harder to spot and identify.[1]

Mandiant recently reported on the use of large language models (LLM) – a subfield of AI – to create more effective phishing lures, writing that malicious operators can leverage AI to create text that reflects natural human speech patterns, generating material for more successful phishing campaigns and initial compromises.[2] Adoption of AI-generated social engineering lures has likely already occurred: AI detectors determined a 71% likelihood that generative AI was used in a DarkGate malspam campaign using phishing emails.[3]

## Malicious Code

Over the past year, reports have surfaced that threat actors are already using the generative AI program ChatGPT to write malicious code by using prompts that trick the chatbot into thinking that it is in developer mode. This malicious use of AI could enhance the capabilities of lesser-skilled actors to quickly write code that performs common malware functions such as port scanning, file enumeration, encrypting files, data upload, etc.

Additionally, threat actors may develop new variants of existing malware, rewrite malware in different languages, or create malware configuration files and command and control protocols, enhancing their existing capabilities without investing the time necessary to develop the improved malware. Reduced timeframes for malware production would increase the number of attacks on the lower end of the malware capability spectrum.[4] However, as it presently stands, some level of technical proficiency is still needed to check code and make corrections, limiting its adversarial use and adoption.

In March 2023, researchers from HYAS Labs demonstrated a proof-of-concept (POC) attack, which they call BlackMamba. The POC uses prompts and queries to exploit LLMs to synthesize a polymorphic keylogger functionality on the fly, dynamically modifying the benign code at runtime without command and control (C2) infrastructure. By eliminating C2 communication and generating new, unique code at runtime, the malware was undetectable by at least one industry-leading endpoint detection and response (EDR) solution, according to HYAS Lab's analysis.[5]

## BlackHat GPT Models

Threat actors can manipulate existing publicly available tools by circumventing safeguards, either directly through jailbreak prompts or through a wrapper connected to an LLM application programming interface (API), though researchers report that the technologies are not yet well-equipped to write malware and that criminal exploitation of AI is not mature. Nonetheless, developers are turning to ChatGPT to help them refine their code or generate base code that can be tweaked or refined later.[6]

For example, in June 2023, WormGPT was released and sold on HackForums, promoting itself as a new, uncensored LLM, free from any limitations posed by OpenAI, created specifically for cybercrime activities. But in an 8 August 2023 interview with Brian Krebs[7] WormGPT's author admitted that WormGPT had guardrails, including "anything related to murders, drug traffic, kidnapping, child porn, ransomwares, financial crime." On 8 August 2023, the author stopped selling WormGPT.

Other threat actors have advertised malicious LLMs on underground forums, such as FraudGPT and WolfGPT, promoting their capability to generate social engineering lures and malicious code.[8] Though advertised as custom built LLMs, the tools tend to route the user's requests to publicly available tools such as ChatGPT by using stolen accounts through a VPN connection and using prompt injection to circumvent content filters and safeguards. While difficult to confirm, external factors such as the speed and cost of development, their removal from Telegram channels shortly after being posted, and lack of verifiable evidence indicate that these tools were likely simply rerouting requests to legitimate LLMs.[9]

## Vulnerability Discovery/AI Fuzzing

Fuzz testing or fuzzing is a legitimate black box software testing technique – one in which results can only be interpreted by inputs and outputs – that finds "implementation bugs using malformed/semi-malformed data injection in an automated fashion," according to the Open Worldwide Application Security Project (OWASP).[10] Newer versions use machine learning to prioritize the text strings and values most likely to cause problems.

Though intended for network defense purposes, fuzz testing tools can be leveraged for nefarious purposes. The financial sector is at particular risk from attackers seeking to conduct fraud and harvest credentials with the enhanced capabilities provided by fuzzing.

For example, though ChatGPT controls refuse prompts that are openly illegal or contravene OpenAI Community Standards, researchers have successfully prompted ChatGPT to make a copy of a whole website code. The experiment was taken further by copying sites of financial institutions and a US credit card operator.[11]

## Prompt Injection Attacks

- NIST defines prompt injections as "An injection vulnerability where malicious prompting can cause unexpected model outputs, causing it to bypass security measures or override the original instructions of the GenAI application."[12]
- **Direct prompt injections,** or jailbreaking, occur "when a malicious user overwrites or reveals the underlying system prompt. This may allow attackers to exploit backend systems by interacting with insecure functions and data stores accessible through the LLM," according to OWASP.[13] Attackers with direct input to the LLM can control it by using jailbreak commands that overwrite a chatbot or other AI tool's guardrails. For example, legitimate LLMs will not provide instructions for committing identity theft, but an LLM manipulated with a suitable jailbreak prompt will relate detailed directions.[14]

Some researchers have stated that direct prompt injections are uniquely dangerous because LLMs are not uniform and that AI developers, such as OpenAI, do not yet completely control their products. OpenAI's website states that though it has "made efforts to make the model refuse inappropriate requests, it will sometimes respond to harmful instructions or exhibit biased behavior."[15]

- **Indirect prompt injection attacks** occur when an attacker inserts an illegitimate prompt within an external source – such as a website or file – that is accepted by an LLM. The malicious prompt hijacks the conversation context and causes the LLM to act as a "confused deputy" for the attacker. Researchers believe this attack vector could permit remote control of the model, persistent compromise, theft of data, and denial of service as detailed in the graphic below, created by Cornell University researchers to depict injection methods and impacts.[16] "Additionally, indirect prompt injections do not need to be human-visible/readable," according to OWASP, "as long as the text is parsed by the LLM."[17]

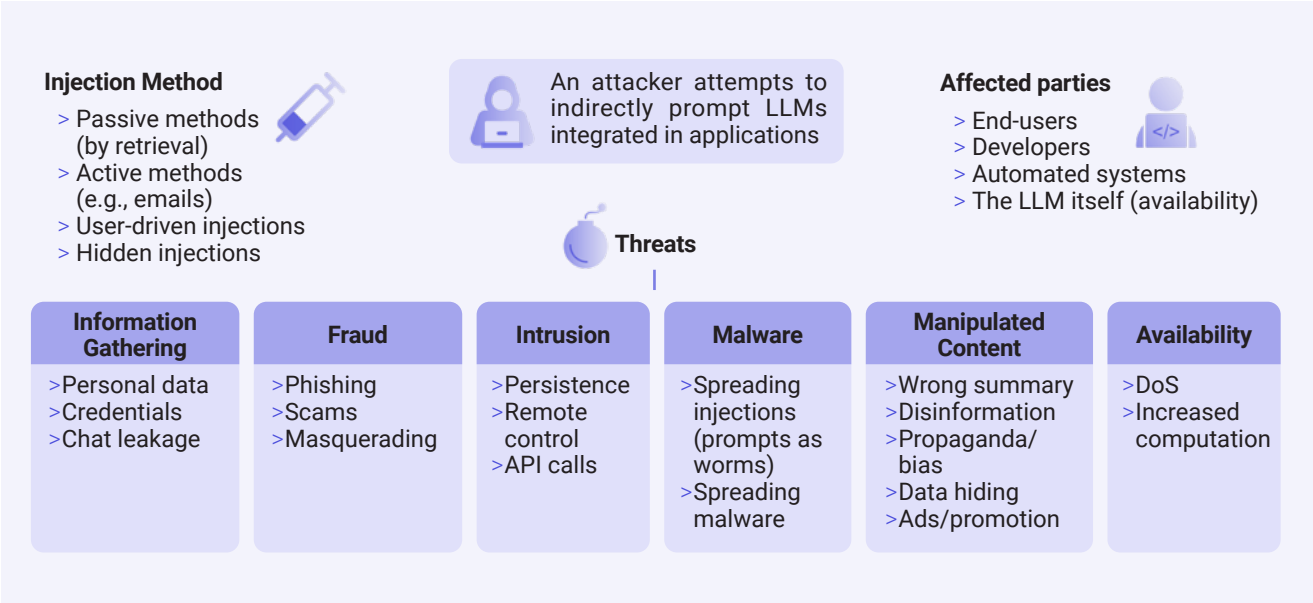



*Figure 2: A high-level overview of new indirect prompt injection threats to LLM-integrated applications, how the prompts can be injected, and who can be targeted by these attacks.*

## AI Poisoning Attack

In an AI poisoning attack, adversaries modify an AI training dataset by injecting it with malicious or misleading data. Those injections alter – or "poison" – the AI output, creating misleading, biased, or incorrect outputs and invalid decision-making processes.[18] By poisoning the training data, the attacker can introduce vulnerabilities, backdoors, or other means of manipulating the behavior of the model. These attacks not only affect the attackers' target, but also undermine the reliability and trustworthiness of AI systems more broadly.

## Deepfakes

Deepfakes are counterfeit videos, audios, and photographs intended to mislead their audience. Audio deepfakes have been used in attacks against corporations and individual consumers.[19] Deepfakes can be used to bypass biometric authentication[20] and to impersonate others over video conferencing.

There are both open-source and vendor solutions that claim to be able to detect deepfakes. However, solutions tend to be domain specific (e.g., just audio, or just text) and are difficult to test against actual

organizational use cases without test datasets consisting of new deepfakes. A further threat is the "adaptive adversary," i.e., one with access to the same solution as the test. This access allows an adversary to craft a deepfake and test it against the solution, incorporating subtle perturbations until it is mistaken for legitimate content. It is recommended that organizations determine where deepfake detection would most benefit them and pursue solutions that meet those needs as best as possible given known limitations, reinforce employees' awareness, and train staff to follow established procedures.

Longer-term solutions include provenance of information (as recommended by the Coalition for Content Provenance and Authenticity[21]), and, where possible, watermarks on newly released video and audio that prohibit their use by generation algorithms.

## Information Operations

The tools to create realistic deepfake videos and fake news articles are readily available – many are open-source – and capable of creating targeted disinformation campaigns designed to deceive and to damage an organization's or individual's reputation. There are deepfake videos of high-profile people, including Volodymyr Zelensky[22] urging Ukrainians to surrender, and two known examples of Russian President Vladimir Putin broadcast on Russian television, one declaring martial law[23] and one deepfake inserted in an interview with President Putin.[24]

### ▸ Disinformation Operation


The greatest threat of disinformation is virality, which is difficult to thwart. Organizations often focus on limiting the extent of the threat actor's intended actions, and can take steps to help prevent disinformation campaigns and mitigate the damages.

- Monitor any content purporting to be from or about senior leaders.
- Watermark legitimate videos to help identify fake videos more quickly.
- Conduct regular training that raises employees' awareness of this threat vector.
- Collaborate with FS-ISAC and others, including large technology platforms, to detect and respond to incidents that suggest the development of disinformation operations.
- Plan an exercise to test detection and response capabilities. Include security operations, forensics, legal, and communications – a multi-prong response across many outlets is required to effectively refute a targeted disinformation attack from a motivated attacker.

## Targeting of AI-Based Solutions

Adversarial targeting of AI-based solutions, including data poisoning, integrity and privacy attacks, attacks against the software supply chain, prompt injection attacks, and other risks vary according to the intent of the adversary and the nature of the targeted systems. Safeguarding against polymorphic attacks requires a multifaceted approach that is conceptually and practically similar to the software development life cycle: both processes must focus on continuous improvement and address security at each stage.

## Data Poisoning

Data poisoning involves the intentional manipulation of training data to influence AI behavior. Successful AI poisoning can cause models to make incorrect or biased decisions or introduce vulnerabilities and backdoors into targeted systems.

> Data poisoning can also be used to protect online content, as illustrated by the Nightshade tool created at the University of Chicago. Nightshade allows creatives to insert data into their content that directs AI models to interpret, for example, images of cats as cows and images of dogs as toasters, which helps prevent the inclusion of uncompensated work in AI models' datasets.[25]

### Integrity and Privacy Attacks

Integrity attacks erode trust in output from AI systems through the compromise of the model's functionality or performance via manipulation of input, the model parameters, or the model's data processing, among other methods. This tampering produces outcomes inconsistent with the solution's intent, which undermines its credibility. Privacy attacks aim to extract information from AI-based solutions – a significant concern as these technologies commonly handle sensitive data. Privacy attacks against LLM applications, for example, can lead to unauthorized access, data leakage, or privacy breaches.

Examples of integrity and privacy attacks include:

- **Unintended data memorization:** Machine learning models can unintentionally memorize training data, such as credit card numbers. Data extraction attacks can obtain such text sequences by, for example, querying the training data in the language model.
- **Membership inference attacks:** Attackers accomplish inference attacks by asking whether a specific data point was used in the training dataset in order to extract private information or details about the dataset. Specific language models are particularly susceptible to inference attacks.
- **Model inversion:** Optimization techniques can be used to find an input that would produce a similar output. By iterating this process, an attacker can gradually reconstruct the original input data.

## Open-Source Compromise

AI models, notably LLMs, increasingly rely on open-source software and public data sources, allowing adversaries to target the components, libraries, or services on which these solutions depend. Attackers with control of update mechanisms for AI applications could push malicious updates to users. Known examples of open-source attacks include:

- The poisoning of publicly available data used to train AI models. Any AI solution trained on the affected dataset could inherit the biases or vulnerabilities introduced through these attacks.
- Inserting malicious code into an open-source library, making dependent solutions vulnerable to exploits enabled through the inserted vulnerability.
- Prompt injections – crafting or manipulating input data in an adversarial attempt to trick AI into taking undesired action – to capitalize on the fact that AI systems are often designed to act autonomously without human intervention. In a well-publicized exploit, a "Do Anything Now" (DAN)[26] attack was used to specifically bypass the safety and moderation features of ChatGPT. This allowed actors to exploit the model and override content moderation features that were designed to ignore or reject prompts that would produce content that is violent, sexual, illegal, unethical, etc.

### ▶ Recommendations, Countermeasures, Defense Mechanisms

- Sanity checks on training data.
- Versions and access controls.
- Adversarial development methods of penetration testing.
- Robust modeling techniques.
- Stress-testing statistical guardrails and detections.
- Defensive programming within application layers.
- Cross-training to bridge the gap between AI/ML and cybersecurity expertise.

## Third-Party Data Sources

> Third-party AI models may be trained on open-source or community-gathered training data. Users need to understand how that data was gathered and if those conditions are similar to the user's use case. For example, a cybersecurity tool's anonymous, cleaned network traffic may reflect "nominal" conditions for customers in general, but may or may not be the same "nominal" conditions in the user's firm.

Users should investigate the data's source, validation techniques, and cybersecurity processes to verify its quality.

> **Sources:** Massive, commonly available datasets (such as Wikipedia and Stack Overflow) typically regulate their information to improve accuracy. But those sites are occasionally subject to interference from malicious users, and their communities can overlook or innocently incorporate incorrect information. As a result, users must pay particular attention to specific areas of concern relating to the desired output.

> **Validation:** Users should understand the process by which the community (or owners) validate and allow the data into the collection, whether the process is fair and equitable, if it contains unintentional bias, and if the source has reasons to exclude certain data that might be needed for a use case.

> **Cybersecurity:** Users should know if and when their vendors scan their datasets for malicious content, specifically OWASP errors, malware, crypto-miners, and other unwanted code. Though more relevant to source code, this approach is applicable to all datasets. Even if the source uses tools to try to limit issues, users should nonetheless scan the output with their own tools.

> **Datasets:** Source datasets tend to evolve over time, and users need to understand the overall security of the dataset so they can be aware of how and when the dataset receives new data or modifies/removes data, and to ensure datasets are not maliciously tampered with (as discussed in the AI Poisoning section). Users need to understand the date or version of the dataset so they know the data that exists in their own models. Users also need to consider what to do when the source modifies/removes data, as specific data cannot be removed incrementally from a model. The only way to ensure the complete removal of data is to retrain the entire model, which is a very resource intensive activity. Further, training data can cause legal liabilities or risk for the user, such as:

- Other firms' proprietary data that could make users liable to infringement claims.
- Licenses for data usage. For example, some firms do not accept certain open-source software licenses, but if source code written under that license is in the training set, then the output is also potentially under the same license.

## Inadvertent Risks

Generative AI poses the additional, inadvertent risks of hallucinations, unintentional biases, data leakage, and anthropomorphizing. Given that these types of risks can include unanticipated consequences, this section is unlikely to be comprehensive.

### Hallucinations

Hallucinations are the phenomenon by which LLMs provide incorrect information presented in a factual manner.

#### ▶ Causes

Hallucinations are not completely understood at this time, and research is ongoing to mitigate their occurrence. It has been shown that language models not only hallucinate but also amplify hallucinations, even those that were designed to alleviate this issue.

It is believed that divergences in the source content (which would often happen with large training datasets) contribute to hallucinations. However, hallucinations still occur when there is little divergence in the dataset. In that case, the phenomenon derives from the way the model is trained, likely:

> An erroneous decoding from the transformer.

> A bias from the historical sequences that the model previously generated.

> A bias generated from the way the model encodes its knowledge in its parameters.

#### ▶ Threats/Risks

> Incorrect information could be used internally within a firm, leading to errors in content

generated for internal use. Equally, incorrect assessments could be provided by an LLM, leading to faulty business decisions.

- Incorrect information could be provided to the customer/consumer, leading to legal, regulatory, and reputational impact.
- Misinformation can be disseminated or promoted on public platforms, leading to reputational damage.
- Vulnerabilities could be introduced into code bases through vulnerable generated code input into wider code repositories for internal tooling and infrastructure.
- Malicious attacks could be instigated, such as package hallucination attacks, in which a threat actor registers a package based on package name hallucinations provided by LLMs, with the intent of infecting a user through use of public repository packages.

### Mitigations

- **Multiagent debate –** Have multiple LLM models debate until a consensus is reached.
- **LLM fact check –** Utilize another LLM to fact check the initial answer.
  - Example: *NeMo Guardrails (NVIDIA)*
- **Validating low-confidence generation –** Use web search results to validate the accuracy corresponding to the low-confidence generation of the model.

According to an active detection and mitigation approach proposed by Cornell University that reduced the hallucinations of the GPT-3.5 (text-davinci-003) model from 47.5% to 14.5%, on average, "first identify the candidates of potential hallucination leveraging the model's logit output values, check their correctness through a validation procedure, mitigate the detected hallucinations, and then continue with the generation process... The detection technique achieves a recall of ~88% and the mitigation technique successfully mitigates 57.6% of the correctly detected hallucinations. Importantly, our mitigation technique does not introduce new hallucinations even in the case of incorrectly detected hallucinations, i.e., false positives." [27]

## Unintentional Biases

AI training datasets can consist of historical data – some LLMs have been trained on all of the text available on the internet – and biases within that data can be learned by the AI model. For example, one image generator consistently portrayed doctors as White males even when prompted explicitly to provide a Black doctor because the majority of the images provided for training contained White doctors.[28]

### Threats/Risks

- Bias in a financial institution's AI system can cause failure to meet regulatory requirements, such as fair lending practices, or respond inappropriately due to the prompts used in the language model or chatbot. This can result in regulatory and reputational risks.

### Mitigations

- **Training of AI models** can include labelled data (excluding often biased characteristics) combined with supervised learning, aimed at reducing unintentional bias.
- **Testing of models** before production should include test suites to elicit any potential biases.

## Training Data Sets

Data collection for training and testing also runs risks, such as intentional or unintentional data leakage, theft, and quality issues.

### Threats/Risks

- Centralized locations for data increase the risk of data leakage. This can be due to, for example, system misconfigurations (such as using default container settings), or through intentional data leakage (e.g., insider access).
- Centralized locations reduce the effort necessary for IP theft.
- The prevalence of AI-generated content in online spaces creates the potential for its inadvertent use as training data, lowering the quality of the input and leading to the over-tuning of models.

### Mitigations

- **Security practices,** such as appropriate access control systems and access control review. Ensure security monitoring is in place.

## Anthropomorphization

People tend to assign human attributes to inanimate objects, which can affect customers' behavior and opinion of financial services companies.

> Many people were outraged at a video showing a robotic dog being kicked to demonstrate its stability,[29] and personality traits such as moodiness and flirtatiousness have been attributed to LLM models.[30]

### Threats/Risks

- Emotional attachment to a chatbot design (especially if the LLM behind the chatbot is coupled with a realistic synthetic voice) or other internet tool. That attachment could cause users to:
  - Provide sensitive/inappropriate information that provokes legal and regulatory issues or data poisoning (if the model learns from inputs).
  - Accept output – including hallucinations – from the tool as more factual/valid than other sources, increasing the risk of hallucinations being transferred from tool to product.
  - Apply the tool to use cases that it is not designed for, leading to higher levels of inaccuracy.
  - Cause individuals to prefer a specific outdated or vulnerable version of the tool, reducing the likelihood of patching and broadening the attack surface.

### Mitigations

- **Training and testing chatbots** to ensure they provide appropriate information in accordance with the organization's communications standards.
- **Disclaimers** to identify when an individual is conversing with AI and not a human.
- **Prohibit responses** to non-professional questions.
- **Sentiment analysis** to ensure that responses are neutral in tone.

## Environmental Concerns

Through the analysis of vast datasets and pattern recognition, AI presents the opportunity to improve energy efficiency, reduce waste, and enhance existing sustainable practices.[31]

Nonetheless, the energy requirements for training and operating AI models are high, with the majority of that energy still coming from non-renewable sources.[32] This increase in energy use directly affects greenhouse gas emissions, aggravating climate change. According to OpenAI researchers, since 2012, the amount of computing power required to train cutting-edge AI models has doubled every 3.4 months.[33] It is estimated that 14% of the world's emissions will result from the Information and Communications Technology (ICT) industry by 2040, largely from ICT infrastructure, particularly data centers and communication networks.[34] Training a single AI system can emit over 250,000 pounds of carbon dioxide, and the use of AI technology across all sectors produces carbon dioxide emissions at a level comparable to that of the aviation industry.[35, 36]

The rapid evolution of the technology and the ever-increasing requirements to maintain operation is likely to further proliferate electronic waste, which often contains hazardous chemicals, including lead, mercury, and cadmium, that can contaminate soil and water supplies and endanger both human health and the environment.[37]

## External Factors
## Legal, Regulatory, and Ethical Threats

Many national government organizations are interested in regulating the development and use of artificial intelligence, and some have existing legal frameworks that apply to AI use, such as those

regarding copyrighted works in training data. Threat actors are unlikely to observe or abide by any regulations. The goal of the legal and regulatory space, in addition to protecting the public interest, is to reflect the ethical values of a society. That said, views on the ethical usage of material or algorithms differ.

### Threats/Risks

- Multi-national organizations adhering to the strictest regulations are at a competitive disadvantage to organizations that meet local, less rigorous regulations.
- The regulatory environment can pose risks to financial institutions (and other organizations) if decisions made in the pre-regulation environment do not meet current regulatory requirements. Depending on the regulation, the required changes could be easily mitigated, or they could require removing or replacing the existing application completely.
- Customers may disapprove of AI usage, such as the use of their electronic conversations to fine-tune large language models. Failure to demonstrate compliance with their wishes could potentially result in legal challenges or reputational damage.

### Mitigations

- **Work closely with government and regulatory bodies** to ensure that the financial organization's needs are represented by new legislation.
- **Adhere internal policies to ethical standards** (such as the *FS-ISAC Responsible AI Principles*) to help mitigate legal, regulatory, and ethical concerns.

## Geopolitical Developments

Geopolitical tensions, national disasters, and global events can destabilize operations for trans-national organizations.

### Threats/Risks

- Geopolitical events can create legal and regulatory risks, as well as operational risks. The causes can range from political events (e.g., uprisings, wars) to climate-related events (e.g., fires, hurricanes).
- Geopolitical tensions between nations competing on AI development are likely to cause additional risk to institutions utilizing AI and AI-adjacent technologies.
- Sanctions, natural disasters, and other restrictions on computing hardware could result in shortages of physical hardware required for training a model, impact storage capabilities for training datasets, and restrict access to cloud-hosted applications. For instance, given that TSMC, the world's largest semiconductor manufacturer, is located in Taiwan, the tensions between China and Taiwan have strong influence over the accessibility of semiconductors and thus the computational power required to further AI development.
- The use of hardware, software, or data provided by global firms presents the threat of legal and regulatory restrictions against pre-built applications, similar to the US removal of Huawei telecommunications infrastructure across critical infrastructure entities.[38] This presents a specific threat for AI models, due to the difficulty in removing the influence of specific data points and subjects.

### Mitigations

- **Draft a disaster recovery plan** relevant to the threat landscape and how resources are being used across geopolitical boundaries.
- **Ensure redundancy methods are implemented appropriately** to mitigate impact from denial of service.
- **Use datasets and third-party components from trusted sources and entities.**

## Group Chair

Benjamin Dynkin, Chair

Hiranmayi Palanki, Vice Chair

## Contributors

Dr Donnie Wendt, *Mastercard*

Monica Maher, *Goldman Sachs*

Hiranmayi Palanki, *American Express*

Benjamin Dynkin, *Wells Fargo*

Elizabeth Geary, *FS-ISAC*

Mike Brizendine, *Goldman Sachs*

Brandon Kelly, *FirstBank*

Mike Silverman, *FS-ISAC*

## References and Resources


1 https://www.fbi.gov/news/speeches/director-wray-s-remarks-to-the-atlanta-commerce-and-press-clubs

2 https://www.mandiant.com/resources/blog/threat-actors-generative-ai-limited

3 https://www.forescout.com/resources/darkgate-loader-malspam-campaign/

4 RBC White Paper CHATGPT – GENERATIVE AI CAPABILITIES AND THREATS

5 https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware

6 https://www.trendmicro.com/vinfo/es/security/news/cybercrime-and-digital-threats/hype-vs-reality-ai-in-the-cybercriminal-underground

7 https://krebsonsecurity.com/2023/08/meet-the-brains-behind-the-malware-friendly-ai-chat-service-wormgpt/

8 https://twitter.com/DailyDarkWeb/status/1684932827911954432

9 https://www.trendmicro.com/vinfo/es/security/news/cybercrime-and-digital-threats/hype-vs-reality-ai-in-the-cybercriminal-underground

10 https://owasp.org/www-community/Fuzzing

11 RBC White Paper CHATGPT – GENERATIVE AI CAPABILITIES AND THREATS

12 https://www.nist.gov/itl/ai-risk-management-framework

13 https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0_1.pdf

14 https://www.popsci.com/technology/prompt-injection-attacks-llms-ai/

15 https://www.netizen.net/news/post/3346/prompt-injection-generative-ais-largest-vulnerability

16 https://arxiv.org/pdf/2302.12173.pdf

17 https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0_1.pdf

18 https://www.cobalt.io/blog/data-poisoning-attacks-a-new-attack-vector-within-ai

19 https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/

20 https://www.wsj.com/articles/i-cloned-myself-with-ai-she-fooled-my-bank-and-my-family-356bd1a3

21 https://c2pa.org/

22 https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

23 https://www.nytimes.com/2023/06/05/world/europe/putin-deep-fake-speech-hackers.html

24 https://www.nbcnews.com/video/putin-quizzed-about-ai-and-body-doubles-by-his-apparent-deep-fake-200210501620

25 https://nightshade.cs.uchicago.edu/whatis.html

26 https://www.fastcompany.com/90845689/chatgpt-dan-jailbreak-violence-reddit-rules

27 https://arxiv.org/pdf/2307.03987.pdf

28 https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it-

29 https://www.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html

30 https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

31 https://insights.grcglobalgroup.com/the-environmental-impact-of-ai/

32 https://www.statista.com/statistics/269811/world-electricity-production-by-energy-source/

33 https://www.scientific-computing.com/analysis-opinion/true-cost-ai-innovation

34 https://www.emerald.com/insight/content/doi/10.1108/JICES-11-2021-0106/full/html

35 https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

36 https://www.nature.com/articles/d41586-018-06610-y

37 https://earth.org/e-waste/

38 https://www.reuters.com/business/media-telecom/us-fcc-bans-equipment-sales-imports-zte-huawei-over-national-security-risk-2022-11-25/



## Contact

fsisac.com
media@fsisac.com