FS-ISAC

Building AI Into Cyber Defense

February 2024

Contents

Executive Summary

While the risks of artificial intelligence to financial services cybersecurity are substantial, so are the rewards. Artificial intelligence (AI) can automate processes, scan and analyze data, and generate reports – among many other capacities – which saves cybersecurity teams time and greatly expands their scope and impact. The purpose of this paper is to highlight the use of AI-based solutions to bolster an enterprise's cyber defenses and risk technologies, specifically the areas and use cases of AI in cybersecurity. As such, it was engineered to be a practical tool for financial services cybersecurity teams assessing the potential of AI solutions in their function.

That said, AI use cases are frequently cross-functional. Some AI-powered solutions have governance, policy, and compliance implications. Others impact strategic planning and operations. Some even facilitate internal communication. Those considerations, as well as the paper's brief analysis of "build vs buy" choices and the high-level technical architecture associated with integrating such solutions, may make FS-ISAC's AI Risk Working Group's analysis pertinent beyond cyber defenses.

However, the goal of this paper is to examine the key considerations and opportunities of AI in the highly regulated risk and security space. The rewards of the opportunities detailed here are considerable even – or especially – to a cyber community highly alert to the risks of AI.

Introduction

AI learns from historical data to predict outcomes, make recommendations, or facilitate decisions. Those capacities have diverse applications; currently, AI is frequently used to predict malicious and benign alerts within networks. As such, firms can significantly benefit from integrating AI into their enterprise cybersecurity and risk management programs.

Further, AI offers operational efficiencies with significant business implications to the financial services industry. Models pretrained on vast datasets can summarize text, answer questions, generate content, and detect similarities between documents, among other tasks. Capabilities like those significantly reduce redundancy and democratize the usage of high-performing tools, advancing research while lowering the barriers to entry. Importantly, many pretrained models can be fine-tuned with custom data, providing a level of adaptability that was previously unimaginable.

The result is a form of AI with a wide range of applications that conducts the most mathematically intensive work during the training stage, simplifying the process of loading and utilizing pretrained models for a wide range of applications. Nonetheless, potential pitfalls and technical considerations, as well as some opportunities and risks, should be considered when incorporating AI into cyber defense.

AI Applications in Cybersecurity and Risk

New developments in AI can be applied to common financial services cybersecurity use cases, including detecting anomalies, creating structure in unstructured data, generating content, and retrieving data. Those use cases are examined below as they apply to domains and functional roles.

Anomaly Detection

Analysts often have to triage a slew of data, both structured and unstructured, to identify outliers or anomalies for the purpose of early incident detection and response. AI systems are trained to find patterns within complex data structures, enabling them to easily identify data points that do not conform to the accepted patterns. Algorithms like DBSCAN, Isolation Forests, Bayesian Networks, and AutoEncoders are all effective at anomaly detection.

Anomaly detection can be applied to a diverse set of cybersecurity and risk domains, as identified in Table 1. However, there are so many ways to apply anomaly detection across enterprises that it can increase the signal to noise ratio (SNR) due to excessive false positives. Fine-tuning the model to filter the noise is recommended, in balance with the supporting team or organization's operational interests.

Creating Content and Structure in Unstructured Data

Threat reports, policies, standards, controls, and industry frameworks often manifest as sprawling 'long-form' text documents that can be time consuming to read and analyze. One of AI's transformative abilities lies in its capacity to help make sense of unstructured data. Generative AI (GenAI) can effortlessly parse and triage long-form text, which simplifies reporting, a vital link between technical intricacies and board-level decision-making. GenAI can also extract structured data or specific fields, enabling further insights to be derived, such as the conversion of information within a threat report into SIEM queries. And the technology can be applied to phishing simulations, reviewing, and actioning reports, with the potential to reduce false positives (too many of which can numb employees to actual phishing threats).

AI can generate a variety of content, as identified in Table 1, for multiple functional roles. Furthermore, AI can reduce the burden of adapting new industry frameworks by employing semantic similarity models to search for the most relevant internal policies and controls. Mapping internal policy and control documentation cohesively yields operational efficiency wins across the board.

Efficient Data Retrieval

As it does with structuring data, AI excels at data retrieval. As noted in Table 1, such solutions have implications for vulnerability management, compliance, and information analysis – even internal communications.

Security organizations spend a considerable amount of time in ad hoc consulting to their companies around security best practices, standards, procedures, etc. GenAI Q&A systems excel at swiftly answering questions, even those posed by analysts lacking experience in querying ingested data using tools like SQL. These systems can convert user descriptions into query language, executing queries to provide prompt and precise answers. AI-based assistants and chatbots addressing similar needs are also becoming increasingly popular.

Use Case Examples

In Table 1 below, common cybersecurity and risk use cases for AI are highlighted along with their domains.

Use Case	Description	Functional Roles	Domain Type
Security Operations Center (SOC) automation	Automate routine tasks such as log analysis, case summarization, threat hunting, and incident response, freeing up SOC teams to focus on more strategic activities.	Threat hunting, SOC analyst, incident/ticket triage, operations analyst	All
Cyber threat detection	Analyze large volumes of cyber threat data in real-time, identify patterns and anomalies, and predict potential future attacks to help analysts quickly detect and respond to emerging threats. For example, analyzing traffic logs to identify anomalous network traffic that deviates from expected behavior.	Threat detection, hunting, threat intelligence and/or incident response	Anomaly detection
Malware detection	Help identify and analyze malicious files.	Operations analyst, threat hunting	Anomaly detection
Insider threat	Finding anomalous access management requests and entitlements to reduce insider threat risk.	Insider threat, threat hunting	Anomaly detection

Fraud detection	Analyze historical and real-time data to create rules and detections to find anomalies relating to identity theft, payment fraud, account take over, money laundering, and check and credit card fraud. For example, freezing a large transaction of funds being transferred to a new account.	Fraud detection, fraud intel, detection engineering, operations analyst	Anomaly detection
Threat hunting	Analyze data from various sources to identify hidden or emerging threats, allowing analysts to proactively hunt for potential threats and vulnerabilities.	Threat hunting, cyber threat intelligence/ analysis	Anomaly detection
Creating training materials	Create cybersecurity training materials, exercises, and other employee awareness content by leveraging GenAI to assess the threat landscape.	Training and awareness	Content generation
Phishing simulations, reviewing and actioning phishing reports	Running spear phishing simulations, identifying true positives vs. false positives in the employee reported phishing email suspects.	Training and awareness, operations analyst, red team/blue team	Content generation
Open-source intelligence analysis	Analyze unstructured data from news articles, social media posts, and other public sources to identify potential threats and vulnerabilities.	Cyber threat intelligence, fusion center, vulnerability management	Creating structure
NLP for data loss prevention	Analyze text data and identify sensitive information (such as personally identifiable information or credit card numbers) that may be at risk of being leaked for data loss prevention programs.	Data governance, protection, data loss prevention analyst, fraud prevention, operations analyst	Creating structure
Keeping up to date with industry standards and policies	Scrape the most recent documentation of known security and privacy frameworks (e.g., NIST, ISO, SOC, GDPR) and compare it to the corporate internal standards to make suggestions for edits and introducing new standards.	Governance, policy, regulatory compliance	Creating structure
Security consulting	Security AI assistants/chatbots to answer ad hoc queries around security best practices or controls and save consulting hours.	Cybersecurity analyst, consultant, advisor, network security	Efficient data retrieval
Vulnerability management and prioritization	Identify and prioritize patch management by locating data from network devices, endpoints, and other infrastructure components.	Vulnerability management	Efficient data retrieval
Risk management/ assessment	Analyze and identify key risk-relevant information in documents, contracts, controls, and policies against compliance and regulatory standards.	Risk management, regulatory compliance	All

Table 1

On the Other Hand: The Risk Considerations of AI

AI systems are vulnerable to a variety of risks, which are examined more thoroughly in the FS-ISAC AI Risk Working Group white papers, *Adversarial AI Frameworks: Taxonomy, Threat Landscape, and Control Frameworks; Responsible AI Principles;* and *Combating Threats and Reducing Risks Posed by AI*. A brief list is provided here to offer perspective to financial services cybersecurity professionals as they seek to balance AI's potential with its risks.

- **Legal, regulatory, and ethical threats:** Given the lack of regulation in a changing legislative landscape, compliance threats and ethical considerations present a unique concern.
- **Unintentional biases:** Faulty and offensive outputs are correlated with data input and the type of learning utilized. Labeled data that excludes often-biased characteristics combined with supervised learning minimizes but does not eliminate the risk of bias.
- **Geopolitical disruption:** Political instability, natural disasters, and other widescale disruptions can threaten the manufacturing and distribution of the technology components and software required to develop and maintain AI models, as well as affect the rules and regulations on use of data for training.
- **Vulnerable centralized databases:** The large and complex datasets required to train AI models create centralized databases that are targets for data poisoning and IP theft attacks. Misconfigurations in cloud environments also present opportunities for significant data leakage.
- **Exposure of sensitive data:** Models that are trained on a massive dataset of public code repositories could generate code that contains sensitive information, such as API keys, passwords, or other credentials. Exposing sensitive data is a concern with most such models.
- **Hallucinations:** A phenomenon in which GenAI gives misleading responses, code, text, or other deviations from facts. Hallucinations are a serious threat because they are difficult to detect and mitigate and can negatively impact the company's reputation. Hallucinations can happen when a large language model (LLM) is trained on data that contains errors or biases, or when it is given incomplete or ambiguous instructions.

Best Practices for Addressing Risk

Defenders should be aware of the current reliability limitations of AI and validate responses for accuracy and trustworthiness before purchasing or implementing AI solutions. Some best practices when leveraging or testing AI include:

- Human-in-the-loop validation that makes AI outputs recommendations, not directly performing actions (applies to both classical and generative AI systems).
- GenAI guardrails that identify harmful responses (applies to generative AI systems).
- Interpretability/explainability analyses using standard explainability strategies (mainly applies to classical AI).

The Confidentiality section of the FS-ISAC paper, *Framework of an Acceptable Use Policy for External Generative AI,* has further information on generative AI-specific risk considerations.

Vendor Solutions vs. Internally Built Solutions: A Capability Perspective

When deciding between a vendor solution and an AI system built internally, it is important to consider the capabilities, customizability, and risks associated with each approach. In this section, we will review these considerations, including the architectural components and strategies that can be used to mitigate implementation risks.

Note that the FS-ISAC AI Risk Working Group has also produced *The Generative AI Vendor Evaluation and Risk Assessment* white paper to help organizations assess and select generative AI vendors while managing associated risks. The white paper's companion workbook simplifies the vendor evaluation process and ensures that financial services institutions make informed decisions when considering generative AI solutions.

The Tradeoff Between Open- and Closed-Source Data

One of the central distinctions between external vendor offerings and custom-built AI solutions lies in the data on which they are trained.

Some external vendors incorporate their customers' data into their training processes, which inserts valuable internal data – such as language, network structure, or other distinctive attributes – into the model. When internal source datasets, such as the specific language embedded in policy documents, are required, a more customized solution may be necessary. However, this proposition raises substantial contractual concerns pertaining to the storage and mining of proprietary information.

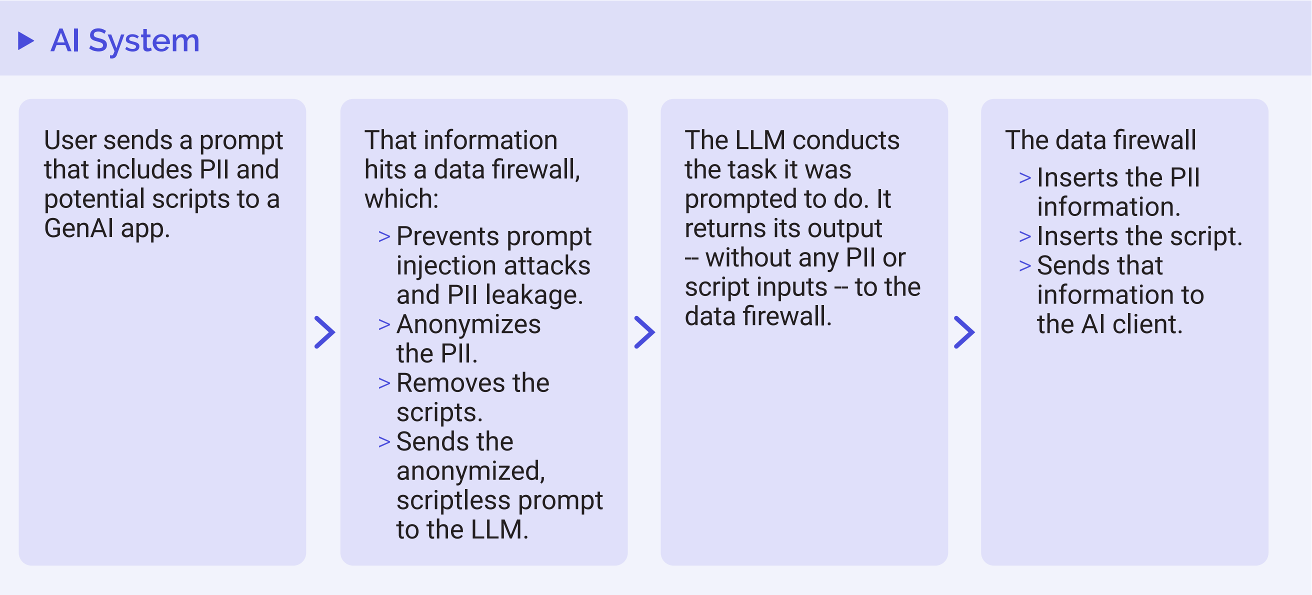
External datasets lack that risk, but they also lack unique internal training data. That may suffice when open-source data aligns with the information the AI system needs in order to learn. Such is the case in some threat intelligence use cases in which identifying trending security topics often relies on processing news articles and social media posts. In any case, it is paramount to meticulously delineate the knowledge base an AI system should comprehend to make an informed decision between a custom or external offering.

Additionally, one must assess the maturity of the solution; it's worth noting that security considerations may be somewhat neglected in general-purpose AI products.

External Offerings: Embedding Learning Systems in the Architecture

Many corporations are using deeply embedded AI products that deploy an 'ensemble of models' to address specific tasks, notably within the domain of SOC automation. These ensembles are initially pretrained on other customers' data (with their consent) and further refined as users interact with the product. The advantage of this approach is in harnessing the power of crowdsourced data and custom data. However, this method risks exposing internal data, a consideration that should not be taken lightly.

Whether the model is integrated as an external offering within or outside of the architecture, firms can always add a data firewall between AI clients and the models (e.g., LLMs) to mitigate the risk of data exfiltration (as shown in the graph below). These firewalls can detect and anonymize personal identifiable information (PII) before it is shared with the model, then de-anonymize it before the output is returned to the AI client. Firewalls can also detect and prevent prompt injection attacks.



▸ AI System

User sends a prompt that includes PII and potential scripts to a GenAI app.

That information hits a data firewall, which:

- Prevents prompt injection attacks and PII leakage.
- Anonymizes the PII.
- Removes the scripts.
- Sends the anonymized, scriptless prompt to the LLM.

The LLM conducts the task it was prompted to do. It returns its output -- without any PII or script inputs -- to the data firewall.

The data firewall

- Inserts the PII information.
- Inserts the script.
- Sends that information to the AI client.

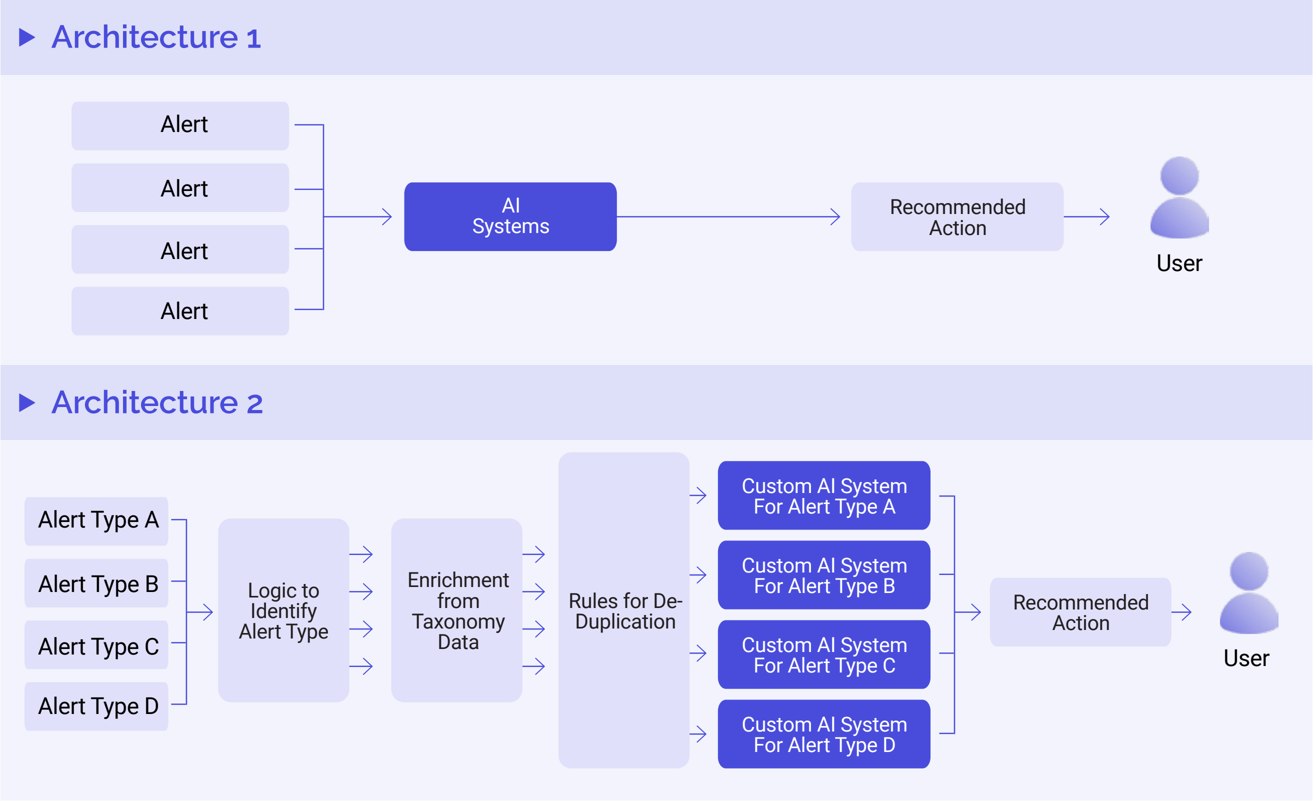
Architectural Design

AI systems, particularly generative AI systems, often manifest as "black boxes" – technology that receives input and produces output with little clarity on how that output was generated. This black box characteristic, far from being a consequence of poor design, arises from the inherent complexity of AI systems.

However, data sensitivity and the impact of outputs can be exceptionally significant in cybersecurity and risk.

The importance of designing AI systems with explainability at the forefront cannot be overstated. Architecture 2 below highlights some of the commonly used explainability methods used by model developers. In addition to using these methods at the model-level, practitioners can make system-level decisions to increase interpretability. As the decision-making process unfolds, one of the pivotal determinants lies in identifying precisely 'where' in the pipeline the AI system is integrated.

Consider, for example, a simple alert integration in which SOC analysts use AI to help manage alerts on the network. There are two proposed AI architectures:



Architecture 1
Alert
Alert
Alert
Alert
AI Systems
Recommended Action
User
Architecture 2
Alert Type A
Alert Type B
Alert Type C
Alert Type D
Logic to Identify Alert Type
Enrichment from Taxonomy Data
Rules for De-Duplication
Custom AI System For Alert Type A
Custom AI System For Alert Type B
Custom AI System For Alert Type C
Custom AI System For Alert Type D
Recommended Action
User

Both of the architectures take in alerts and give back some recommended action. When a recommendation is wrong, however, it is important to determine how the system arrived at its solution. In Architecture 1, the AI system is trained on all types of alerts, and it is up to the learned patterns to determine the best action. These learned patterns are immensely complex and difficult to look at "under the hood" – the definition of black box.

In Architecture 2, the AI system is integrated further right in the pipeline. A number of steps take place on the alerts before they are fed into the AI system that is customized to what was determined in the previous steps. These models in system 2 need not be as large and complex as system 1, since they are more purpose driven. Identifying the factors that led to a given recommendation is a great deal easier as well.

External offerings may be much more complex than the architectures highlighted above. Users should ask questions on the design so they can effectively and confidently take action on data-driven outputs.

Conclusion

The defense and cybersecurity landscape is witnessing an "AI epidemic" as enterprises and vendors alike look for ways to embrace AI. Throughout this paper, we have explored several pivotal points in this landscape: We have defined what "AI" is, unearthed common applications in security and risk domains, scrutinized the risks associated with intricate black box models, delved into architectural strategies to enhance explainability, and discussed the considerations to make when deciding between vendor or custom-built solutions.

Ultimately, when evaluating the suitability of an AI solution, whether it be a vendor's or an internal solution, a firm's primary assessment revolves around the tradeoff between the value AI adds and the risk of incorrect and unexplainable outputs.

Fortunately, these risks can be mitigated through architectural design choices for internal products or by judiciously selecting vendors. In integrating AI into defense systems, it is paramount to understand these issue points, balancing the appeal of AI with the imperative of safeguarding against unexplainable missteps.

Group Chair

Benjamin Dynkin

Contributors

Jamie Weiss, *American Express*

Monica Maher, *Goldman Sachs*

Benjamin Dynkin, *Wells Fargo*

Joseph Lavelle, *PNC*

Kapil Pruthi, *TIAA-CREF*

Mike Silverman, *FS-ISAC*

References and Resources

Table 2: Explainability Methods

Name	Applicable Models	Technical Reference
Partial dependence plots	Model agnostic (classical ML)	Interpretable machine learning (1)
Shapley Additive Explanations (SHAP)	Model agnostic (classical and generative ML)	Implementation reference (2) SHAP For transformers (3)
Individual Condition Expectations (ICE)	Model agnostic (classical ML)	ICE paper (4)
Attention maps	Transformer MODELS	Attention map article (5)

1. https://christophm.github.io/interpretable-ml-book/pdp.html

2. https://shap.readthedocs.io/en/latest/

3. https://aclanthology.org/2021.hackashop-1.3.pdf

4. https://arxiv.org/pdf/1309.6392.pdf

5. https://ai.plainenglish.io/visualizing-attention-in-vision-transformer-c871908d86de

Contact

fsisac.com

media@fsisac.com