Treasury's Report on AI (Part 2) —Managing AI-Specific Cybersecurity Risks in the Financial Sector

July 3, 2024

This is the second post in our two-part Debevoise Data Blog series covering the U.S. Treasury Department's report on Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector (the "Report").

In Part 1, we addressed the Report's coverage of the state of AI regulation and best practices recommendations for AI risk management and governance. In Part 2, we review the Report's assessment of AI-enhanced cybersecurity risks, as well as the risks of attacks against AI systems, and offer guidance on how financial institutions can respond to both types of risks.

Cyberattacks Enhanced with AI

As we have previously discussed, companies face unique cybersecurity issues related to AI. In its Report, the Treasury Department outlined several ways malicious actors can use AI to enhance existing cyberattacks against financial institutions.

- **AI-enhanced social engineering**. Social engineering is a well-known tactic that involves the use of non-technical means to execute cyber-attacks – most often by manipulating victims to perform actions in furtherance of the attackers' goals. As we have previously described, AI can allow attackers to create more targeted and more persuasive social engineering attacks with less resources – attackers can now create deepfake audio and video in real time. These attacks may be particularly effective when the attacker understands the finance industry's emphasis on prompt, low-friction transactions and can persuade victims and intermediaries to act before taking the time to second-guess the attacker's requests.

  What firms can do: Financial institutions should consider conducting additional training that will help employees identify the hallmarks of suspicious requests, including instructions to send money or confidential information to a new account, a sense of urgency, a requirement for secrecy, or directions to forgo controls. Financial institutions should also consider requiring employees to follow additional

verification protocols for wire transfers above a threshold to new accounts and when sending confidential materials to any location outside the company, including by phone or video conference.

- **Malware/Code generation and automated discovery of vulnerabilities**. Commercially available AI systems may enhance the ability of attackers to create and deploy malware, as well as to discover vulnerabilities in networks and applications used by financial institutions. According to the Treasury Department's report, current security programs that rely on signature-based detection systems may not be sufficient to detect AI-modified malware. And some reports suggest that general-purpose AI can already outperform specialized non-AI systems in detecting vulnerabilities.

  What firms can do: Financial institutions should consider using cybersecurity tools that incorporate AI in order to supplement security programs otherwise based on detection of known artifacts or indicators of compromise. Those tools may be essential in detecting new and evolving threat tactics that are themselves the product of AI. Simultaneously, firms should also consider adopting security principles that account for the increase in the exploitation of vulnerabilities, such as Zero Trust Architecture.

- **Disinformation using AI**. The Treasury Department notes that threat actors may be able to use AI to increase the reach and persuasiveness of false information. This can include spreading damaging information about a company, harming the company's reputation, and disrupting daily business operations, which may serve other goals of threat actors.

  What firms can do: Financial institutions should consider implementing procedures related to verification and authentication, such as digital signatures and asymmetric encryption, and communicating to their stakeholders that information from the companies will always be verified with those predetermined methods. Additionally, firms should consider constantly monitoring for disinformation, so that they can quickly respond to any disinformation before it spreads. Creating a protocol for identifying and taking down false information, and practicing that protocol through tabletop exercises should also be considered.

Cyberattacks on a Company's AI Systems

In addition to summarizing ways that attackers can use AI offensively to cause harm, the Report also describes several ways that cyberattacks can be conducted against the AI systems used by financial institutions.

- **Data poisoning**. Threat actors may be able to purposefully insert information into data used by AI systems that will result in the AI systems producing unexpected or undesirable outputs. Threat actors can use this kind of data poisoning tactic to target AI tools being used to detect and prevent malicious activity, for example, by making an AI tool unable to distinguish between fraudulent and legitimate transactions.

  What firms can do: Financial institutions should consider implementing procedures to protect their data sources, such as limiting access to high-risk AI models and data sets, and enhancing logging and audit capabilities.

- **Data leakage during inference**. Threat actors may be able to cause an AI system to generate an output containing confidential or other sensitive information. This includes, for instance, personal data, financial data, material non-public corporate data, and other content used either to train the underlying models or as reference data incorporated in AI system responses.

  What firms can do: Financial institutions should consider implementing procedures that ensure that information walls and permissions are properly applied to data that are made available to the AI systems. The procedures may include testing the effectiveness of the controls through external testing and implementing risk-based controls to detect and prevent attempted data leakage.

- **Evasion**. A threat actor who understands the relationship between inputs and outputs for a specific AI system may be able to cause an AI system to produce results that a human overseer would not have approved. For instance, an attacker who is applying for a loan through an AI system may manipulate how they present their data in order to receive approval from the AI system.

  What firms can do: Financial institutions should consider enhancing their AI models with training designed to identify evasion attacks, such as adversarial training. Stress testing may be used in conjunction with adversarial training to ensure that AI models can identify evasion attacks with reasonable accuracy. Human oversight may also be used by financial institutions until a decision is made that AI systems can be trusted to make stand-alone decisions.

- **Model extraction**. A threat actor may be able to create a functionally equivalent model of an AI system – effectively stealing the underlying software and IP – through repeated interactions with the system and close analysis of the relationship between the system's inputs and outputs. This can not only devalue the software itself, but it may also allow the threat actor more opportunities to find ways to examine the copy to discover how the original model could be vulnerable.

  What firms can do: Financial institutions should consider implementing mitigation measures, such as session-based limitations to restrict the amount of information that an actor can receive at a given time, as well as controls that restrict the types or content of data that is given to a user. Financial institutions should also consider conducting penetration tests to evaluate the effectiveness of the mitigation measures.

Key Takeaways

Treasury's report did not break new ground in its discussion of the AI-enabled threats that confront financial institutions. But it does provide a helpful inventory of risks for firms to consider when evaluating both their AI controls and their cybersecurity programs. In addition to the considerations above, firms may want to consider the following takeaways based on the Report:

- **Secure the supply chain of AI systems**. Financial institutions may want to consider how they can go about monitoring and assessing the development and deployment of their AI systems throughout their life cycle, both for hardware and software components. For procured AI systems or services, consider conducting due diligence specifically on the extent to which third-party vendors have tested and hardened their AI systems.

- **Implement secure design of AI systems**. Financial institutions may want to implement controls to ensure security risks are contemplated in the design and development of AI systems. Specifically, financial institutions may want to consider gating questions for developers and business owners related to technical risks, trade-offs, and potential vulnerabilities of AI systems they are contemplating for development or procurement. Firms may also want to document consideration of the holistic impact of an AI model were it to be misused or compromised.

- **Ensure secure deployment of AI systems**. Firms should consider controls that will support secure deployment and in-production use of AI systems. Firms may want to have defined cybersecurity expectations for AI systems that are appropriate to the

risks of the system, such as having appropriate access controls, a segregated environment that holds golden-copies of code, as well as verification requirements. Financial institutions also may wish to maintain earlier versions or alternative versions of AI models that could be implemented if the newest version is misused or compromised.

- **Adopt cybersecurity guidance and frameworks to meet regulatory expectations and stay current**. As threats evolve, so do industry standards and best practices, which then inform regulators' views of the adequacy of a firm's technical controls. Firms may already be considering how to update their cybersecurity programs in response to NIST's release of version 2.0 of its Cybersecurity Framework. In addition, financial institutions may want to consider how their controls align with NIST's Artificial Intelligence Risk Management Framework, which the Report specifically references, as well as with other applicable guidance, such as the joint guidance published by the U.S. Cybersecurity and Infrastructure Security Agency and the UK National Cyber Security Centre.

To subscribe to the Data Blog, please click here.

The Debevoise Artificial Intelligence Regulatory Tracker ("DART") is now available for clients to help them quickly assess and comply with their current and anticipated AI-related legal obligations, including municipal, state, federal, and international requirements.

The cover art used in this blog post was generated by Microsoft Copilot.

* * *

Please do not hesitate to contact us with any questions.

Avi Gesser
Partner, New York
+1 212 909 6577
agesser@debevoise.com

Erez Liebermann
Partner, New York
+1 212 909 6224
eliebermann@debevoise.com

Matt Kelly
Counsel, New York
+1 212 909 6990
makelly@debevoise.com

Jackie Dorward
Associate, New York
+1 212 909 7406
jmdorward@debevoise.com

Kyungjin Kim
Associate, New York
+1 212 909 6551
kkim@debevoise.com

Joshua A. Goland
Law Clerk, New York
+1 212 909 6420
jagoland@debevoise.com