Risk of AI Abuse by Corporate Insiders Presents Challenges for Compliance Departments

February 21, 2024

We recently highlighted the need for companies to manage risks associated with the adoption of AI technology, including the malicious use of real-time deepfakes (*i.e.*, AI-generated audio or video that impersonates a real person). In this article, we address three AI-related insider risks that warrant special attention by corporate compliance departments (*i.e.*, insider deepfakes, barrier evasion, and model manipulation) and present possible ways to mitigate them.

Insider Deepfakes. The ability of rogue employees to use AI tools to create very realistic forged documents, as well as deepfakes, poses new dangers for companies. Many compliance systems are designed to require specific approvals for certain employee actions. For example, actions such as significant payments to new vendors, changes to bank accounts for existing payees, business gifts, and reimbursements for work-related expenses, often require certain internal approvals. Just as external threat actors can use modern technology tools to circumvent these controls, the ability to create fake documents, audio or video will make it much easier for insiders to fabricate compliance with these measures. And while insiders are more likely to be discovered and held responsible for their actions, they also are more likely to have privileged knowledge of security processes and workflows, providing greater opportunities to undermine company procedures.

Some of the defensive practices that we have recommended for external deepfake threats apply equally to insider threats, such as requiring dual authorization for high-risk transactions. In addition, insider deepfake risks can be addressed in a company's AI and cybersecurity tabletops, as well as Incident Response Plans. Compliance departments may also consider staying abreast of developments in deepfake detection, such as Intel's recently announced real-time detection software.

But perhaps the most effective way to combat the risks of insider deepfakes is training. Compliance can teach employees that AI technology can now create very convincing fake documents, audio, and videos, and it can be done in real time. Therefore, employees can be sensitized to the fact that any authorization of an unusual expense that is

provided in a document, audio, or video could be fraudulent, especially if the action being approved has one of the following hallmarks: (a) it involves the transfer of large sums of money or highly sensitive information, (b) it does not follow normal protocols, or (c) it has an element of urgency. Training can specifically note that employees will not face any adverse action for following company verification protocols when presented with such an authorization, even if the authorization seemingly came from a company executive.

Information Barrier Evasion. Corporate compliance often focuses on preventing improper access to sensitive information. Compliance departments thus apply robust controls in their information technology environments, often by erecting strict "walls" on who can access which information. These restrictions protect against the impermissible disclosure of sensitive information such as material nonpublic information ("MNPI") and trade secrets.

With AI, gaps in these walls have become easier to discover. Corporate chatbots, for example, can be given access to internal corporate data that employees access by "chatting" with the AI system. The underlying data often includes employee communications, corporate policies and procedures, and swaths of unstructured corporate data. A chatbot given access to sensitive data might divulge it in response to a routine prompt, without the employee even knowing about that possibility.

These capabilities could allow a rogue employee to deliberately extract MNPI or other confidential walled-off information using a chatbot. Such employees might try to cover their tracks by making it appear that they are engaging in a routine interaction with the chatbot, when in fact they are seeking to push the system into revealing information it is meant to keep secret.

One defense is to ensure that existing information walls and permissions are properly applied to the data made available to their AI systems. Implementing zero-trust architecture, including "least-privilege" practice that limits access to employees with genuine need, can also protect sensitive data from AI exploitation. Effectiveness testing, such as through red-teaming (i.e., engaging a team of experts to probe systems for vulnerabilities), can help assess AI information controls, both before and periodically after deployment. Depending on the sensitivity of data exposed to a particular tool, companies might consider implementing risk-based controls to detect and prevent attempted misuse, including automated prompt monitoring and escalation.

Model Manipulation. Lastly, AI systems are being increasingly relied upon to drive vital business processes such as sales and investment models. Rogue employees might

try to tamper with these systems by, for example, manipulating algorithms to artificially boost their performance metrics.

Compliance departments may therefore want to consider limiting the ability of a single employee to alter their AI systems. The maker-checker process, for example, requires that a "checker" approve any changes that a "maker" seeks to implement. This dual authorization greatly improves the ability to detect improper changes, as well as inadvertent errors, made to internal systems.

Additionally, compliance departments might consider auditing the content of their highest-risk AI systems and use version control to detect when changes have been made. With such an audit trail, coupled with detailed access logs, any unauthorized changes can be quickly detected, remediated, and investigated.

Conclusion. The risk of insider threats is too often lost in the discussion of new technologies. AI tools, for all their promise, provide evolving ways for rogue employees to subvert data controls for their own personal benefit. Companies may wish to consider these heightened risks in their AI strategies and include risk mitigation as they plan their control strategy.

Key Takeaways.

- Consider updating trainings for detecting and responding to suspect situations.
- Consider staying abreast of developments in deepfake use and detection.
- Consider using stronger authentication measures, such as biometric authentication and encrypted digital signatures, in order to protect against deepfakes in certain circumstances.
- Consider applying existing information walls and permissions to new AI systems.
- Consider implementing zero-trust architecture for certain high-risk systems.
- Consider implementing risk-based controls to detect and prevent attempted misuse of their tools (e.g., chatbots with access to MNPI), including automated monitoring and escalation.
- Consider implementing dual authorization for any changes in AI systems, which will help prevent model manipulation.

- Consider auditing and version control of information used by high-risk AI systems to better detect and remediate any improper changes to those systems.

To subscribe to the Data Blog, please click here.

The Debevoise Data Portal is an online suite of tools that help our clients quickly assess their federal, state, and international breach notification and substantive cybersecurity obligations. Please contact us at dataportal@debevoise.com for more information.

The cover art used in this blog post was generated by DALL-E.

* * *

Please do not hesitate to contact us with any questions.

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

Douglas S. Zolkind
Partner, New York
+1 212 909 6804
dzolkind@debevoise.com

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

Sarah Wolf
Counsel, New York
+1 212 909 6334
swolf@debevoise.com

Scott Woods
Associate, New York
+1 212 909 6859
sjwoods@debevoise.com

Karen Joo
Law Clerk, New York
+1 212 909 6528
hjoo@debevoise.com