The Value of AI Incident Response Plans and Tabletop Exercises

April 27, 2022

Today, it is widely accepted that most large organizations benefit from maintaining a written cybersecurity incident response plan ("CIRP") to guide their responses to cyberattacks. For businesses that have invested heavily in artificial intelligence ("AI"), the risks of AI-related incidents and the value of implementing an AI incident response plan ("AIRP") to help mitigate the impact of AI incidents are often underestimated.

The Value of CIRPs and Tabletops for Cybersecurity

Cybersecurity programs used to be focused primarily on preventing attacks, but that has changed as companies and regulators have accepted that successful attacks will happen even with great cyber defenses. A key element of a strong cyber program is resilience—the ability of the company to quickly and effectively respond to a successful cyberattack.

Regulators have come to recognize that maintaining a CIRP is one of the best ways for an organization to improve its cyber resilience. Indeed, many cybersecurity regulations, including Part 500.16 of the New York Department of Financial Services' ("NY DFS") Part 500 Cybersecurity Requirements for Financial Services, the FTC's strengthened Safeguards Rule, and the National Association of Insurance Commissioners Insurance Data Security Model Law, require companies to maintain written CIRPs as part of their efforts to reduce the impact of cyberattacks. Regulators also encourage companies to test their CIRPs though tabletop exercises. For example, the June 30, 2021 NY DFS Ransomware Guidance provides that CIRPs "should be tested, and the testing should include senior leadership—decision makers such as the CEO should not be testing the incident response plan for the first time during a ransomware incident."

Incident response plans are helpful because things move very quickly in most cyber and AI incidents. For leaders to have the right information to make timely decisions, multiple stakeholders need to work in parallel on many different tasks. A well-crafted, tested incident response plan provides details regarding who is responsible for which tasks and who has the authority to make which decisions. For example, in the first few

hours of a typical ransomware attack, a strong CIRP will provide a checklist of critical tasks and who is responsible for each of the following:

- deciding whether to proactively take down a portion of the network to stop the spread of the malware;
- protecting the back-up systems and testing whether they are available;
- determining whether the company's email system is safe to use or whether an alternative communication system is required;
- escalating the incident to Legal, Executive Management, and/or the Board;
- assessing the nature and volume of any data that have been encrypted or stolen;
- preserving evidence and protecting privilege;
- deciding whether to engage external support and, if so, contacting and briefing external counsel, a cyber forensic firm, a crisis communications firm, a ransomware negotiator, etc.;
- drafting internal and external communications;
- engaging with the FBI or other government agencies;
- assessing insurance coverage and making any required insurer notifications;
- determining the identity of the threat actor and whether they are subject to the U.S. Department of the Treasury's Office of Foreign Assets Control ("OFAC") designations or other sanctions;
- scheduling regular meetings and determining who attends which meetings;
- assessing any regulatory or contractual breach notification obligations with short deadlines and drafting those notifications; and
- assigning one or more points of contact for internal and external inquiries.

Knowing in advance of an incident (1) what are the most important tasks for the first 24 hours; (2) who is responsible for each of those tasks, as well as who is the back-up in case that person in not available; and (3) how each team member conveys information

to the decision makers, allows for quick effective responses and is often one of the reasons why a small cyber incident doesn't become a large one.

By contrast, organizations that spend the initial few days of an incident figuring out what needs to be done and who should do it, often miss opportunities to contain the attack, and may also make a difficult situation worse by making inaccurate or confusing communications to employees, customers, investors, regulators, and the public.

The Value of AIRPs and Tabletops for AI Incidents

As more companies implement AI for their core business functions, the risks of AI incidents increase. Examples of AI incidents include the following:

- Public complaints of bias or unfair treatment of a protected class resulting from an AI tool used for healthcare, lending, or insurance underwriting, such as news reports that Optum's healthcare algorithm discriminated against black patients, which prompted an investigation by NY DFS.

- Failure to sufficiently disclose that AI is being used for certain tasks or decisions, like in BlueCrest, where investors were not aware that an algorithmic trading application was managing a substantial portion of their funds, leading to an SEC enforcement action.

- Failure of AI tools that are used for core business functions, such as in the case of Zillow, where a house-pricing algorithm performed poorly in light of new circumstances arising from the pandemic, resulting in hundreds of millions of dollars in losses and several shareholder lawsuits.

As is the case for cybersecurity, having a written, detailed, and tested AIRP will lead to more effective responses. For example, if an internal whistleblower or a job candidate accuses a company of using an AI-hiring tool that is biased against a certain category of candidates, it would be helpful for the company to have a written AIRP that provides a checklist of critical tasks and who is responsible for each, such as:

- assessing whether the AI system is currently in use, and if so, the extent of its use and what manual and automated alternatives exist for carrying out those tasks;

- determining if the AI system is feeding data into, or otherwise interacting with, critical AI or other IT systems;

- estimating the financial impact of stopping the use of the AI system;
- deciding whether to proactively stop using the AI system while an investigation is conducted;
- collecting and reviewing logs from the AI system;
- reviewing comments and complaints from users about the AI system;
- escalating the incident to Legal, Executive Management, and the Board;
- assessing the nature and volume of past decisions that have been made by the AI system;
- examining training and testing inputs and outputs of the system, along with any results from ongoing monitoring;
- preserving evidence and protecting privilege;
- deciding whether to engage external support and if so, contacting and briefing external counsel, an AI auditing firm, a crisis communications firm, etc.;
- drafting internal and external communications;
- assessing risks from civil litigation and regulatory action;
- assessing any regulatory or contractual notification requirements drafting those notices;
- examining any statements made about the AI system to consumers, users, or the public;
- assessing insurance coverage and making any required insurer notifications;
- scheduling regular meetings and determining who attends which meetings; and
- assigning one or more points of contact for internal and external inquiries.

In recognition of the value of AIRPs, the recently published guidance on AI bias from the National Institute of Standards and Technology ("NIST") (Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, March 15, 2022), included a

recommendation that companies "detail any plans related to incident response for such [AI] systems, in the event that any significant risks do materialize during deployment."

The European Commission's Draft EU AI Act, does not specifically require an AIRP, but it does include mandatory AI incident reporting to regulators. Under the European Commission's draft, providers would be required to report any "serious incidents" involving high-risk AI systems within 15 days of becoming aware of the incident. Having an AIRP would clearly be helpful to companies in meeting that potential deadline.

Like CIRPs, once an AIRP is drafted, it is very helpful to test its effectiveness through a mock tabletop exercise to ensure:

- the key tasks and decisions are covered;
- the right people are assigned to those tasks and decisions;
- communications, escalations, and approvals are properly addressed;
- external resources are identified, so that the company is not scrambling to identify and retain the right outside advisers during an actual incident; and
- difficult decisions (such as whether to immediately discontinue the use of an AI system that is under scrutiny) are thought through, so that executives are not addressing these complex issues for the first time during an actual incident.

Conclusion

Companies have dramatically reduced the risks and impacts of cyberattacks by being well prepared to respond. Having a written incident response plan that has been tested through a tabletop exercise is a key aspect of that preparation. As companies expand their use of AI, their risk of having an AI-related incident increases, and the benefits of having an AIRP become more evident.

* * *

To subscribe to our Data Blog, please click here

NEW YORK


Avi Gesser
agesser@debevoise.com


Anna R. Gressel
argressel@debevoise.com


Corey Jeremy Goldstein
cjgoldst@debevoise.com


Michael R. Roberts
mrroberts@debevoise.com


Erik Rubinstein
erubinst@debevoise.com