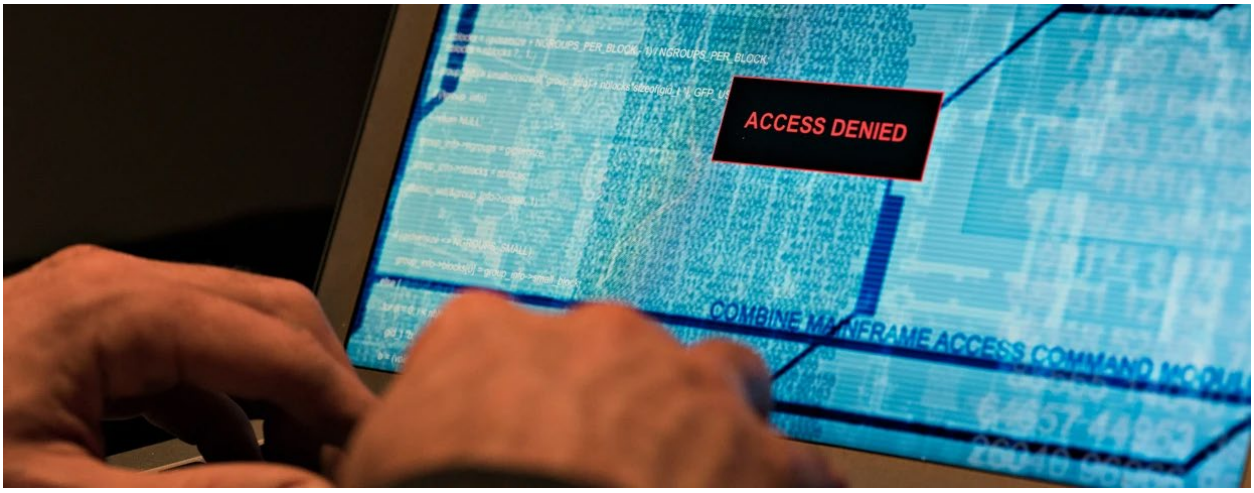
Bloomberg Law®

Groups Using AI Tools Should Prep for Cybersecurity, Legal Risks

July 31, 2025

Kenneth Rashbaum
Barton



Cybersecurity concerns are often lost in the discussion around artificial intelligence even though AI-related security events can create devastating economic and even human danger.

AI tools are ubiquitous, with users who work in organizations ranging from financial services and health care to academia, e-commerce and manufacturing. But AI models can be hacked just like any other form of data because of user errors, vulnerabilities in the model, and so-called “zero-day” exploits (an exploit so new that there are few defenses against it and zero days to respond with a solution).

One key security weakness is shadow, in which users share information with AI models that haven’t been verified and authorized by their organization’s IT department. Writer James Coker estimated that 46% of Gen Z and 43% of Millennials say they have shared information with AI tools without telling their employers.

Hackers can access sensitive or proprietary information from poorly protected systems and then sell it or use it to create deep fakes and believable phishing links. Certain AI models can create their own malicious code and even write an exploit with detailed instructions on how to use it. Other security risks include:

- **Prompt injections:** Hackers embed commands into the prompt to bypass guardrails and reveal sensitive information.
- **Theft of AI source code:** This can be processed and distributed by cybercriminals to train malicious tools.
- **Data poisoning:** This type of manipulation can cause supply chain forecasting to be erroneous, financial projections to be inaccurate, and health procedures to malfunction.
- **Software Vulnerabilities:** AI can learn from them and create zero-day exploits. Because AI models are constantly evolving, there are undoubtedly additional risks that haven't been widely reported and many more in development. A security-first culture with regard to AI acquisition and deployment is therefore critical.

Organizations should consider how that culture is achieved and whether there are other technical, administrative, and legal security controls that can be implemented now. This may include looking at:

- **Access controls:** Limits, with appropriate security controls (including encrypting the training data and multi-factor authentication to access it), on who can access the training data to reduce the risk of cybercriminals injecting malicious code or manipulating the data.
- **Vetting and validating:** Forensic review of the training data to ascertain the presence of corrupted or malicious data and remove it.
- **Oversight:** Continuous monitoring of the model in operation and auditing for anomalies.
- **Adversarial training:** Introducing "adversarial examples" into the training data set so the model can learn to recognize and resist poisoned data.
- **Adherence:** To the National Institute of Standards and Technology AI Risk Management framework to improve resistance to data manipulation, in which adversaries, through "input attacks," can craft patterns of changes to a target that will fool the AI system into making a mistake.
- **Updating of Security Protocols:** To address cybersecurity controls specific to AI incidents and training of the workforce on the updated protocols.

Legal Protections

As we have previously discussed, AI models are constantly evolving. They will continue to evolve beyond the point where years-old template contracts—that can be used for anything a procurement department seeks to purchase or license—can provide adequate protection for customer or vendor. In addition, these old-school contracts don't provide the basis for a sustainable business relationship.

Data share and license agreements between AI providers and customers should comprise cybersecurity representations, warranties, and security appendices that include:

- Representations and warranties regarding the technical controls set out above
- A notice and acceptance clause that permits the customer to "sandbox" (test drive) the model to review its training data and code for corrupted or malicious data
- A clear definition of a reportable security incident and inclusion of such incidents in clauses on performance metrics (i.e., "exploit," "security incident," "malicious code," "data poisoning," and "data manipulation")
- Well-defined rights of the customer to audit the provider with regard to adherence to the contract's terms
- Carve out of security incidents as described above from the limitations of liability clauses

While AI risks can never be entirely eliminated, using these technical and administrative controls in the provider-customer agreement and cultivating a relationship in which there is an open line of communication regarding new cybersecurity threats can make a significant difference. Mitigation of these threats can be effective and provide customers and providers with a level of confidence in deployment of existing and developing AI models.
