Grant Thornton

Control AI cybersecurity risks

Contents

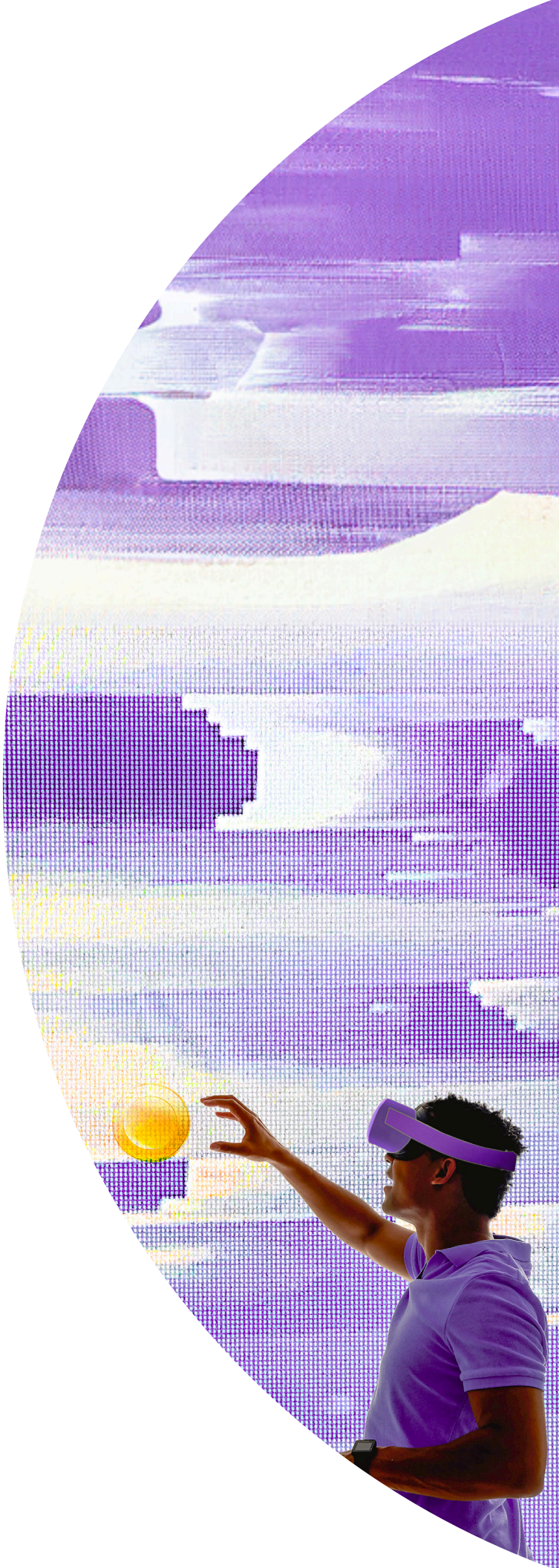
The age of artificial intelligence is not on the horizon.

It's here. The use of artificial intelligence (AI) continues to spread with a speed that's staggering, as it grows even more every day. Companies worldwide have adopted and implemented AI, in solutions that are reshaping industries through improved efficiency, productivity and decision making.

However, the meteoric rise of AI can overshadow some valid concerns around security and privacy.

Many organizations have integrated AI into their business processes more quickly than they have updated security strategies and protocols. This has created complex fractures in outdated programs that leave vulnerable gaps of risk exposure. Your risk, technology and cybersecurity leaders must find, understand and mitigate these exposures.

AI-driven solutions can provide powerful capabilities — for you, and for malicious actors who gain access, data and control from your solutions. Now, many cybersecurity professionals are trying to catch up to rapid AI adoption. But they must go further, to implement forward-thinking strategies that anticipate potential risks, mitigate them and proactively define the security narrative.



Industry use cases for AI

The recent surge in the use and adoption of generative AI (GenAI) technology like ChatGPT has been extraordinary. GenAI has gained visibility due to its free availability, surprising superiority to traditional chatbots, and its rapid integration with search engines like Bing and Google. This powerful AI technology can accomplish a variety of tasks to make our lives more efficient and connected.

In recent months, one case has experienced a meteoric rise in popularity: Content creation. Journalists, bloggers, students and office employees now use GenAI to brainstorm ideas, draft initial versions of articles or papers and polish up their writing. In the education sector, teachers and tutors use GenAI to curate study materials, generate practice questions and simplify complex topics. In many industries, AI-driven language processing can:

- Automate tedious or repetitive work like basic Q&A support and help desk ticket generation
- Enable real-time responses around the clock, without incurring additional human resource cost
- Convert conversation (or "unstructured text") into insights for more informed decision making
- Break down language barriers to facilitate worldwide support and collaboration
- Help write or debug code, as well as optimize and document existing code — this has some people predicting the end of software development as we know it

In recent years, machine learning and deep learning AI have permeated diverse industries to address complex challenges and enable innovative solutions. These solutions have the potential to transform and disrupt how we work, create and consume. The following use cases illustrate some of the ways AI is applied in various industries and domains.

Healthcare

AI systems can analyze ultrasounds, X-rays, MRIs and other medical images to provide insights and inform diagnosis and treatment. For example, the FDA has approved an AI system that can analyze scans to identify signs of stroke and help physicians quickly make treatment decisions. In the height of the COVID-19 pandemic, AI played a significant role in helping to determine responses and treatments. One system uses deep learning algorithms to analyze medical images for features that are indicative of COVID-19. AI has also been used to predict potential mutations of the SARS-CoV-2 virus which causes COVID-19.

Law enforcement

Facial recognition technology can be used to identify individuals by analyzing their facial features and matching them against a database of known faces. This can have many applications such as identifying suspects in criminal investigations or verifying individuals at border crossings or airports, but it also comes with an explicit risk of misuse by governments or companies that might target or discriminate against certain individuals.

Media and entertainment

Image generation technology is not new, but AI has accelerated and drastically improved the quality and speed of creating and augmenting images from various sources. In the past year, revolutionary GenAI image generation tools have emerged that can turn words and pictures into stunning art. Recent advancements also enable users to send audio data through a GenAI model that creates music, sounds or voices.

Automotive and geolocation

Geo-tagging and real-time mapping, which involve the identification, live tracking, and analysis of objects or locations in the real world, are a rapidly growing set of AI use cases. This capability has significant applications in fields like urban planning, disaster management, environmental monitoring and autonomous vehicles. By accurately identifying and interpreting the physical environment, AI-based systems can provide insights and actionable information for decision-makers, ultimately leading to safer and more efficient operations.

Business intelligence and analytics

AI technologies often improve efficiency with intelligent automation, and decision-making processes with reasoned and insightful analysis. In the financial sector, companies have already explored using machine learning for automated investing, portfolio diversification and retirement planning, optimizing returns based on market trends and investor preferences. Credit bureaus now employ AI to analyze non-traditional data points for more precise creditworthiness predictions. AI's ability to identify anomalies and trends from large data volumes is driving robust fraud detection in banking and retail.

Human resources

The integration of GenAI in HR has transformed talent management and workforce planning. AI tools facilitate more accuracy in automated résumé screening and candidate ranking, reducing the time spent on manual reviews and improving the likelihood of identifying optimal candidates. However, there are legitimate concerns surrounding the widespread use of AI. The primary concern is that potential bias in AI algorithms may drive discriminatory hiring practices or unfair evaluations. There is also a fear of overreliance on technology, which may result in the devaluation of human intuition and judgement in the hiring and HR management process.

Cybersecurity

In the realm of cybersecurity, AI has emerged as a game-changer with numerous applications that are designed to safeguard our digital world.

- **Code generation and review:** AI code generation can help developers create code more efficiently and accurately, while constantly reviewing the code for vulnerabilities, inefficiencies or faulty logic. This enhances the overall quality and security of software while reducing the time spent on remediating insecure code farther down the software development lifecycle.
- **Network intrusion detection systems:** These systems can use AI machine learning and reasoning to identify and establish a baseline for a network's behavior, raising alerts for anomalies or threats. They can even counteract the threats automatically, significantly reducing adversary dwell time.
- **Phishing and spam detection:** This helps protect users and organizations from unwanted and malicious messages. AI natural language processing and machine learning can identify and block harmful emails. Even large email providers, like Gmail, use AI to analyze email content and sender data, filter out spam, detect sophisticated phishing attempts, and continuously refine their detection algorithms. Browsers and security solutions can also identify phishing websites, analyzing URLs and web content to alert users or block the sites.
- **Malware analysis:** AI is widely used in malware analysis to monitor system behavior and detect anomalous activity that could indicate a malware infection. AI technologies can also help generate and test malware samples in safe environments, to proactively identify vulnerabilities and develop countermeasures. AI can also help automate and improve the speed and accuracy of incident response to malware attacks. Advanced solutions can automatically quarantine infected systems, identify and block or filter the source of attacks, and provide real-time updates and mitigation strategies to security teams.

AI technologies can enhance cybersecurity tools, and help you detect and respond to threats more effectively, but the same technologies can also be used by attackers.

The age of AI has significantly reduced the time, money, and expertise necessary to conduct cybersecurity attacks… So, you must be vigilant and proactive in anticipating new risks and preparing their cybersecurity programs.

Cybersecurity risks in AI

As AI tools become more prevalent and sophisticated, they can be used to pose significant cybersecurity risks. Malware can now use AI techniques to evade traditional antivirus software, so traditional approaches and countermeasures might not mitigate new risks like:

- Incorrect or biased outputs
- Vulnerabilities in AI-generated code
- Copyright or licensing violations
- Reputation or brand impacts
- Outsourcing or losing human oversight for decisions
- Compliance violations

As companies adapt their business strategies for new AI capabilities, they must also adapt their risk mitigation strategies. Cybersecurity and data privacy are an essential part of mitigating AI risks.

Implementing AI technologies in your organization can introduce five primary cybersecurity risks.

1. Data breaches and misuse

Data breaches pose a significant cybersecurity risk for AI platforms that store and process vast amounts of confidential or sensitive data like personally identifiable information, financial data and health records.

Several risk factors can contribute to data breaches in AI platforms. Internally, AI instances that process and analyze data can be vulnerable due to weak security protocols, insufficient encryption, lack of adequate monitoring, lax access controls and internal threats. Externally, AI solutions and platforms can be vulnerable to various security risks and can be targets for data theft, especially if data used to interact with these platforms is logged or stored.

The risk of misuse and data loss has increased due to the unfettered availability of GenAI projects like GPT-4 or PaLM2, along with other open-source variants. The risk is especially high for IT, engineering, development and even security staff, who may want to use GenAI to expedite their daily tasks or simply experiment with new tech. They can inadvertently feed sensitive data through browser extensions, APIs or directly to the GenAI platform. Without an enterprise-sanctioned solution, some may use their personal accounts, potentially committing their companies to terms of use that may not be acceptable from the privacy and risk perspective.

2. Adversarial attacks

Adversarial attacks manipulate input data to cause errors or misclassification, bypassing security measures and controlling the decision-making process of AI systems. There are several forms of adversarial attacks, and two of the most common types are evasion attacks and model extraction attacks.

Evasion attacks try to design inputs that evade detection by the AI system's defenses and allow attackers to achieve their goals (like bypassing security measures or generating false results). Since the inputs appear to be legitimate to the AI system, these attacks might produce outputs that are incorrect or unexpected without triggering any detection or alerts. Model extraction attacks try to steal a trained AI model from an organization to use it for malicious purposes. Some applications are particularly vulnerable to these attacks. The impacts of adversarial attacks vary by use case and industry, but can include:

- Errors or misclassifications in the output for medical diagnostics, where adversarial attacks can misdiagnose cases and potentially cause improper treatment. In the context of automated vehicles, such attacks might incorrectly interpret traffic signs and cause accidents
- Decision-making manipulations that could coerce a system into divulging sensitive information or performing unauthorized actions

3. Malware and ransomware

Malware and ransomware have plagued IT systems for years, and even AI platforms can also be subject to these attacks. In fact, AI lowers the cost of malware generation, so attackers can deploy new variants of malware quicker, cheaper and with less skill. The risks for any solution include:

- Disruption of services, caused by encrypting data or overloading networks to prevent legitimate access
- Hijacking resources to use for crypto mining or a botnet attack
- Exploiting publicly available AI platforms to pivot into your network and cause harm

4. Vulnerabilities in AI infrastructure

Like any software, AI solutions rely on components of software, hardware and networking that can be targeted by attackers. In addition to traditional attack vectors, AI can be targeted through cloud-based AI services, graphic processing units (GPUs) and tensor processing units (TPUs).

GPUs and TPUs are specialized processors designed to accelerate AI workloads, and they can introduce new attack vectors. Design flaws in processors and other hardware can affect a range of products. For instance, the Row hammer (or rowhammer) flaw affects dynamic random access memory chips in smartphones and other devices. Attackers can use a "Flip Feng Shui" technique on the Row hammer flaw to manipulate memory deduplication in virtualized environments, compromising the target's cryptographic keys or other sensitive data.
AI solutions are also built upon and integrated with other components that can fall victim to more traditional attacks. Compromises to the software stack can trigger denial of service, gain unauthorized access to sensitive data or pivot into your internal network.

5. Model poisoning

Adversarial attacks target AI models or systems in their production environment, but model poisoning attacks target AI models in a development or testing environment. In model poisoning, attackers introduce malicious data into the training data to influence the output — sometimes creating a significant deviation of behavior from the AI model.

For example, after a successful model poison attack, an AI model may produce incorrect or biased predictions, leading to inaccurate or unfair decision making. Some organizations are investing in training closed large language model (LLM) AI to solve specific problems with their internal or specialized data. These applications can be subject to serious damage from model poisoning attacks without proper security controls and measures in place.

Model poisoning attacks can be challenging to detect, because the poisoned data can be innocuous to the human eye. Detection is also complicated for AI solutions that include open-source or other external components, as most solutions do.

Privacy regulations and compliance

AI solutions can introduce unique cybersecurity and data privacy risks. To address those risks, organizations, developers, users and policymakers must understand some key considerations.

Regulatory considerations for AI use

As AI use and adoption grows, regulations will emerge to help ensure ethical use, data protection and data privacy.

Currently, one of the most comprehensive regulatory frameworks for data protection is the EU's General Data Protection Regulation (GDPR). It governs the collection, processing and storage of personal data. For example, Article 35 of the GDPR requires organizations to perform a Data Privacy Impact Assessment (DPIA) for certain types of processing activities, particularly when new technology like AI is involved. You must be sure that you can comply with the requirements of the DPIA before initiating the processing activities. It could be difficult for you to give regulators an adequate DPIA without a thorough understanding of how an AI model works.

Many other jurisdictions are also considering or have implemented regulations to mitigate AI bias. These regulations typically require transparency in AI-driven decision making, auditing of AI systems for bias, and penalties for companies that fail to comply. One such regulation comes from The New York City Department of Consumer and Worker Protection (DCWP) and is set to begin enforcement in 2023.

Current regulations have varying requirements, but overall, these nascent attempts to regulate AI use are focused on key principles and avoid being too prescriptive. You should assess their current compliance requirements — and the key principles being addressed — to verify that their use of AI can continue to comply.

Privacy concerns from the use of AI in business operations

The use of Personally Identifiable Information (PII) to train AI models has given privacy and security professional a reasonable cause for concern. By incorporating PII into the training process, developers risk creating models that inadvertently reveal sensitive information about individuals or groups. As AI models become more powerful and adaptable, they might also learn to extract sensitive information from users in the course of conversations. A failure to sufficiently protect PII could lead to privacy breaches, scams, phishing and other social engineering attacks.

To mitigate these risks, consider a range of factors and potential issues in AI technology:

1. Loss of sensitive information

One of the most pressing concerns is the potential exposure of sensitive information that end users enter in conversational AI systems. While this information may seem harmless on its own, it can be combined with other data points to create detailed profiles of individuals, potentially jeopardizing their privacy.

2. Model explainability

Many advanced AI models are so complex that even their developers might see them as "black boxes." That makes it challenging for organizations to explain the models and their results to regulators. In heavily regulated industries like finance and healthcare, regulators often require clear explanations of a model's outputs and decisioning processes. A lack of explainability can lead to undiagnosed errors and unidentified process improvements, along with more serious issues like undetected biases and ethical implications. It also obscures who or what is responsible when things go wrong. Some of these risks can be mitigated by adopting ethical AI development principles, promoting transparency, and improving user awareness and vigilance. However, mitigations must continue to evolve as AI use continues to expand in scale and complexity.

3. Data sharing and third-party access

AI platforms can involve collaboration between multiple parties, or use third-party tools and services. This increases the risk of unauthorized access or misuse of personal data, especially when data is shared across jurisdictions with different privacy regulations.

4. Data retention and deletion

Some AI solutions store data for extended periods so that they can continue referencing, analyzing and comparing it as part of informing their machine learning, predictive and other capabilities. This long-term data storage increases the risk of unauthorized access or misuse. The context and complexity of AI solutions can also make it challenging to ensure that data is deleted when it is no longer needed or when individuals exercise their rights to request deletion.

5. Inference of sensitive information

Increasingly sophisticated and pervasive AI capabilities can connect and infer sensitive information about users based on inputs that seem innocuous on their own. For instance, inferences could combine inputs to identify political beliefs, sexual orientations or health conditions, posing a layer of risk that is hard to identify without a comprehensive analysis across potential data connections. Even when data is pseudonymized, AI might be able to use advanced pattern recognition or combine datasets to re-identify individuals without permission.

6. Surveillance and profiling

AI technologies like facial recognition and social media monitoring can enable invasive surveillance and profiling of individuals that endangers rights to privacy, anonymity and autonomy.

Mitigation strategies

Many organizations have begun to mitigate risks from generative AI by enhancing privacy solutions and monitoring AI models. To mitigate the cybersecurity risks in new AI solutions, you should review and update your existing cybersecurity program. Programs must include appropriate security measures and technologies to safeguard data and systems from inadvertent mistakes and malicious attempts.

Consider the following aspects when building security and privacy practices in the age of AI:

Policies and procedures

Review and amend existing policies and procedures to define the necessary AI-specific security requirements, designate roles to oversee the AI operations and ensure implementation of the security guidelines.

Threat modeling

Conduct threat modeling exercises to help identify potential security threats to AI systems and assess their impact. Some common threats to models include data breaches, unauthorized access to systems and data, adversarial attacks, and AI model bias. When you model threats and impacts, you can identify a structured approach with proactive measures to mitigate risks.

Consider the following activities as part of your threat modeling:

1. Criticality
Document the business functions and objectives of each AI-driven solution, and how they relate to the criticality of your organization's operations. This helps you to establish a baseline for criticality, making controls commensurate with the criticality of the AI application and determining the thoroughness of the threat model.

2. Connections
Identify the AI platforms, solutions, components, technologies and hardware, including the data inputs, processing algorithms, and output results. This will assist in identifying the logic, critical processing paths and core execution flow of the AI that will feed into the threat model and help edify the organization on the AI application.

3. Boundaries
Define system boundaries by creating a high-level architecture diagram, including components like data storage, processing, user access and communication channels. This will help you understand the AI application's data and activity footprint, threat actors and dependencies.

4. Data characteristics
Define the flows, classifications and sensitivity for the data that the AI technology will use and output. This will help determine the controls and restrictions that will apply to data flows, as you might need to pseudonymize, anonymize or prohibit certain types of data.

5. Threats
Identify potential threats for your business and technologies, like data breaches, adversarial attacks and model manipulation.

6. Impacts
Assess the potential impacts of identified threats, and assign a risk level based on vulnerability, exploitability and potential damage.

7. Mitigation
Develop and implement mitigation strategies and countermeasures to combat the identified threats, including technical measures like encryption, access controls or robustness testing, along with non-technical measures like employee training, policies or third-party audits.

8. Adaptation
Review and update the threat model on an ongoing basis as new threats emerge or as the system evolves.

Data governance

Use effective data governance to help ensure that data is properly classified, protected and managed throughout its life cycle. Governance can include:

1. Roles and responsibilities
Establish policies with roles and responsibilities for data governing, along with requirements for documenting data provenance, handling, maintenance and disposal.

2. Data quality assessments
Regular data quality assessments help identify and remove potentially malicious data from training datasets in a timely manner.

3. Data validation
Data validation techniques like hashing can help ensure that training data is valid and consistent.

4. Identity and access management
Identity and access management policies can help define who has access to training data, with access controls to help prevent unauthorized modifications.

5. Acceptable data use
Acceptable data use policies can outline what data can be used and how it can be used. Each data classification (like public, internal, confidential, or PII) should include its uses and restrictions pertaining to AI technologies. Policies should also include procedures for users to follow if they find a restricted data type in an AI solution or training set.

Implement secure data management and governance practices to help prevent model poisoning attacks, protect data security, maintain data hygiene and ensure accurate outputs.

Access control

To control access to your AI infrastructure, including your data and models, establish identity and access management policies with technical controls like authentication and authorization mechanisms.

To define the policies and controls you need, consider:

1. Who should have access to what AI systems, data or functionality?
2. How and when should access be re-evaluated, and by whom?
3. What type of logging, reporting and alerts should be in place?
4. If we use AI with access to real data that may contain PII or other sensitive information, what access controls do we need, especially as related to the data annotation process?

Reassess and update policies and technical controls periodically, to align with the evolving AI landscape and emerging threat types, ensuring that your security posture remains robust and adaptable.

Encryption and steganography

Encryption is a technique that can help protect the confidentiality and integrity of AI training data, source code and models. You might need to encrypt input data or training data, in-transit and at-rest, depending on the source. Encryption and version control can also help mitigate the risk of unauthorized changes to AI source code. Source code encryption is especially important when AI solutions can make decisions with potentially significant impacts.

To protect and track AI models or training data, you can use steganographic techniques like:

- **Watermarking** that inserts a digital signature into a file, or the output of an AI solution, to identify when a proprietary AI is being used to generate an output.
- **Radioactive data** that makes a slight modification to a file, or training data, to identify when an organization is using the training data. For instance, radioactive data can help you protect your public data against unauthorized use in the training of AI models.

End-point security, or user and entity behavior analytics

End points (like laptops, workstations and mobile devices) act as primary gateways for accessing and interacting with AI systems. Historically, they have been a principal attack vector for malicious actors seeking to exploit vulnerabilities. With AI-augmented attacks on the horizon, end-point devices warrant special consideration as part of the potential attack surface.

User entity and behavior analytics (UEBA) enabled end-point security solutions can help detect early signs of AI misuse and abuse by malicious actors. UEBA is known for its capability to detect suspicious activity by using an observed baseline of behavior, rather than a set of predefined patterns or rules. As a result, it offers a more effective solution than rule-based or supervised-learning security tools. UEBA employs advanced techniques like unsupervised machine learning and deep learning to detect new patterns and abnormal behaviors, providing a more dynamic and adaptive approach to identifying potential security threats.

Vulnerability management

AI systems can be vulnerable at many levels, like the infrastructure running the AI, the components used to build the AI or the coded logic of the AI itself. These vulnerabilities can pose significant risks to the security, privacy and integrity of the AI systems, and you need to address them through appropriate measures like robust security protocols, testing and validation procedures, and ongoing monitoring and maintenance.

You should regularly apply software updates and patching to keep all software and firmware components of the AI infrastructure up to date. You should conduct regular assessments of AI infrastructure components, including hardware, software and data, to identify and remediate vulnerabilities in a timely manner.

Conduct periodic penetration tests on the AI solutions and functionality. This ensures that patches on infrastructure are working as intended, access controls are operating effectively and there is no exploitable logic within the AI itself.

Security awareness

With the advent of any new technology, you need to ensure that executives, developers, system engineers, users and others understand the appropriate uses and the risks.

Board members and executives need to know:

- Privacy and data protection regulations that are applicable to their use
- Ethical implications of AI technologies (like potential biases, discriminatory and unintended consequences)
- Legal and regulatory compliance requirements affected by use of AI, like intellectual property, liability and accountability

Users need to know:

- How they can and cannot use AI
- What data is permitted to be used with AI
- Which knowledge base that AI is using, and procedures for reporting incidents

System engineers need to know:

- Guidelines for designing, building and integrating AI systems in a secure and compliant way
- Processes, tollgates, security reviews and approvals required for AI solutions
- Resources and knowledge bases available for AI solutions

Developers need to know:

- Security coding standards to which they must adhere
- Approved repositories and libraries
- Processes, tollgates, security reviews and approvals required for AI solutions
- Resources and knowledge bases available for AI solutions

Every role needs security training that includes specific responsibilities along with these core topics:

- Your security responsibilities
- The processes you are required to follow
- Resources you can use to learn more, or whom you can contact

To improve resilience against threats and safeguard sensitive data, you need to foster a culture of security awareness. You also need to regularly update security training materials to keep pace with the rapidly evolving threat landscape and emerging techniques.

Immediate actions for security teams

Your security team needs to select and design the right mitigation strategies to define a clear roadmap, with prioritized milestones and timelines for execution.

The first stop on this journey should be to define rules of the road. For instance, that can include rules that guide the secure and responsible use of GenAI — review and augment existing policies related to acceptable use, business email and data sharing with third parties.

To take the next steps on your roadmap, consider some important security questions:

- **Do we have the right policies, standards and procedures in place to tackle AI-related security and privacy risks?**
  Acceptable use policies typically address how users should use computing resources like networks, systems and software. These policies are meant to ensure that people use these resources in a responsible, ethical and legal manner. Explicitly include GenAI or other AI technologies in these policies, alongside existing use provisions for websites, social media, email and communications, to emphasize the potential risks involved.

  Revisit your third-party data sharing policy, which typically outlines the types of data that can be shared, the parties that can receive the data, the purposes for which the data can be used, and the methods to ensure security and privacy of the data. These policies should include either the prohibition or limitation on the types of data that may be used during conversations and interactions with AI-driven solutions.

  Finally, conduct security training and awareness campaigns to address the risks associated with the use of GenAI and other AI technologies, including appropriate uses, how to identify and respond to potential security and data breaches, and whom to contact in the event of an incident.

- **Do we need new policies, standards and procedures to cover any gaps in the existing practices or emerging domains?**
  When creating new security policies addressing AI usage in the workplace, it's important to consider several factors. Consider the following to help ensure the security and privacy of employees, customers and others whose confidential information you might hold:

  – ***Who can use AI solutions, for what purpose and under what circumstances?***
  Some organizations might have strict URL filtering protocols to block access to GenAI and other AI solution portals, along with social media sites. However, most organizations rely on less draconian "acceptable use" policies that provide guidelines about how and when to use these tools to make work more efficient. Similarly, some organizations may only allow access for specific people or even specific endpoints. Web browsing from corporate servers has always been frowned upon, and in many circumstances disabled, but access to the internet and appropriate websites has generally been allowed for employees and management without many restrictions.

  Consider providing explicit guidance for IT personnel, engineers and developers on using or integrating AI tools into existing applications or software, to help ensure adherence to the appropriate data sharing policies and standards.

– ***Do we have a well-defined, documented and socialized data classification policy?***
You should have a documented data classification policy and use data loss prevention tools to enforce the policy. The use of data in AI technology, or for the purposes of training technology, should be included in these policies. You should explicitly train and remind users to not use any PII, proprietary information, patent information, source code or financial information in interactions with external AI systems.

– ***How do data privacy, data retention and terms of service affect the adoption and use of GenAI?***
Evaluate when and how to let people use external GenAI technology in your enterprise, in accordance with your enterprise policies, culture and values.

Consider how the GenAI creators approach and communicate their commitment to privacy, their data retention policies, and clauses included in the terms of service (ToS). The GenAI data collection, processing and sharing practices should be clearly outlined and published so that you can answer:

- *Is input data pseudonymized, anonymized or encrypted?*
- *Does the GenAI provider comply with published privacy-related policies and regulations?*
- *Does the GenAI provider mention their data retention and destruction processes?*
- *Does the GenAI provider publish and release any third-party audit reports, certifications or attestations demonstrating their commitment to security?*
- *In the ToS, are there clauses that may identify any potential legal, regulatory or compliance risks and do they include limitations on data usage, sharing and intellectual property rights?*

– ***How do you properly integrate AI into existing or newly developed applications or software?***
The development or integration of AI should start with a threat modeling exercise. Any significant changes to existing software should merit a risk assessment. In this way, you can document the risks associated with the software, calculate impact and implement commensurate controls to reduce the risks to acceptable levels.

– ***Whom do employees contact and report when AI results are questionable or expose undesirable information?***
Establish a clear accountability for the AI policies, processes, technology and implementation at all levels of leadership. Help ensure that employees are aware of the process for reporting erroneous outputs, or suspicious activity, to appropriate individuals for follow up and resolution.

– ***What are the opt-out policies for AI, and how do people use them?***
Opt-out policies give users, customers or employees a set of choices and mechanisms to decline or withdraw their consent for solutions to collect, process or share their data. In the EU, the GDPR emphasizes an opt-in approach where businesses must obtain explicit, informed and freely given consent from users. The GDPR also recognizes the "right to object," which lets individuals opt out of certain data processing activities. The U.S. does not have a unified federal privacy regulation, which means that opt-out policies are sector-specific or state-specific (as with the California Consumer Privacy Act).

As you plan software integration and data storage, consider the potential need to implement opt-in or opt-out policies in the future. If users ask to review, modify or delete their data in the future, you need to ensure that you can find and manage all relevant data — and that you can show regulators you have that capability.

See Appendix A for a sample of a GenAI-specific security policy. The policy considers many of the above points, to help stay ahead of AI adoption by employees and potential integrations into existing tools and applications.

Start with understanding

The rapid growth and adoption of AI technologies have brought about significant advancements and improvements in various industries, streamlining processes and enhancing productivity. However, alongside these benefits come concerns related to privacy, security, misuse and ethical considerations.

As AI continues to reshape our world, it is imperative for businesses, governments, regulatory agencies and individuals to collaborate on developing responsible AI practices and regulations that address these concerns. We need to focus on transparency, fairness and ethical use, while mitigating potential risks, to harness the power that AI can have to create a more efficient, innovative and inclusive future.

Appendix: Example GenAI security policy

Purpose

The purpose of this policy is to establish a commitment to the responsible use of Generative AI (GenAI), which includes various technologies and large language models (LLMs). Our organization acknowledges the potential risks and benefits associated with GenAI, and aims to ensure that any implementation of GenAI aligns with the organization's risk posture and risk appetite while also remaining consistent with the organization's values and mission.

Acceptable use policy

Employees may not disclose the organization's confidential, restricted, internal or proprietary information when interacting with GenAI technology, directly or through a third-party application. The following guidelines should be followed when using GenAI systems. Employees should be trained on specific policies and the use of the GenAI system.

Acceptable use of GenAI:

1. Employee use of GenAI should be directly related to their job duties and responsibilities.

2. GenAI should be used for legal and ethical purposes and not violate any applicable laws, regulations or policies.

3. Any data used with GenAI must be owned or authorized for use by the employee and must not infringe on any intellectual property rights.

4. Employees should ensure the accuracy and integrity of any result generated by GenAI before utilizing them in their internal or external communications or any organization intellectual property.

Unacceptable use of GenAI:

1. Using GenAI for any illegal or unethical purposes, including but not limited to discrimination, harassment or fraud

2. Using GenAI to generate false or misleading information

3. Using GenAI to access unauthorized data or systems

4. Using GenAI to create content that infringes on intellectual property rights

5. Using GenAI in connection with taking any internal or external training or examination

GenAI use and implementation guidelines:

1. Identify and document systems, infrastructure, technology and processes using or accessing GenAI technologies.

2. Conduct periodic risk assessments to identify and rectify potential risks related to the use of GenAI technologies.

3. Evaluate the potential effect of GenAI on all stakeholders, including customers, employees and partners.

4. Implement security controls to mitigate risks associated with GenAI, including but not limited to third-party access, authorized access, data protection and incident response.

5. Organize educational programs to inform employees and other users about GenAI technology, its risks, and practices for engagement, including reminders that GenAI-generated responses are not from humans and that their input may be reused by the system in the future.

6. Establish clear channels for employees to report policy violations or questionable results from GenAI technology.

7. Evaluate GenAI technology, infrastructure and business processes when embedding or integrating GenAI into software or applications:
   - Map the potential attack surface by identifying vulnerabilities in GenAI input and develop mitigating controls
   - Implement measures to protect sensitive information, prevent unauthorized access and ensure business continuity
   - Implement logging and archiving capability to record all GenAI usage in accordance with legal and regulatory requirements and document any limitations in data collection

- Log and track actions taken by GenAI and users, flagging anomalies for further analysis
- Conduct explainability research to understand and interpret the decision-making process to increase user confidence and adoption
- Implement security measures and access controls to protect data and systems accessed by GenAI applications
- Grant GenAI applications and integrations only the necessary access, data and permissions to complete their tasks
- Restrict GenAI access to data relevant to the task at hand, ensuring that it does not have access to unrelated or sensitive information
- Assume that GenAI may be compromised through prompt injection and plan mitigating controls
- Assess risks to intellectual property, terms and conditions, opt-out mechanisms, data retention policies, end user licenses and click through agreements
- Ensure that GenAI-generated output is accurate, reliable and free from false or fabricated answers or citations

Contacts

Derek Han
Principal and Leader,
Cybersecurity and Privacy
Advisory Services
T +1 312 602 8940

Maxim Kovalsky
Managing Director,
Cybersecurity and Privacy
Advisory Services
T +1 646 354 0463

Caesar Sedek
Managing Director,
Cybersecurity and Privacy
Advisory Services
T +1 213 596 3465

Grant Thornton