Security and risk considerations:
Safeguarding validity, privacy and transparency

As the utilization of generative AI continues to expand, understanding the associated security implications becomes crucial for businesses across industries.

There are 5 key risk considerations for the accounting profession, both for internal and client use:

1 Privacy

a. Understand the information you're ingesting into AI and its sensitivity. Public tools such as ChatGPT often rely on user input for its own training.

b. "De-identify" (sanitize) personal information before ingesting it into both internal AI systems and public tools. Personal identifiable information is anything that can be used to identify an individual person.

c. Have a conversation with the provider of your enterprise AI platform to exercise control over usage, retention, and disposal of information stored on the platform.

2 Transparency

a. Engage legal counsel to identify any laws that may apply which require disclosure use of AI.

b. If a client's confidential information will be shared with a third party vendor, including AI, provide written disclosure to and obtain specific consent from the client in the appropriate format before the confidential information is shared.

c. If client information is used to train the AI model, then consider explicitly stating such in your disclosure to and consent from the client.

d. If disclosure is not specifically required, consider still disclosing a firm's use of AI in the engagement letter. Include AI in the disclosure if it is anticipated AI tools will be used to render the agreed upon services.

e. An addendum to an engagement letter or other written consent from the client can be created if it is later determined that AI tools will be used.

CPA.com™

3 Avoiding Bias

The way in which we ask questions are often biased, so sometimes an AI can respond to a question in a way which confirms our bias. We need to be thoughtful about both inputs and outputs (responses).

4 Human Review

a. Accounting professionals have a responsibility to monitor answers that are coming out of generative AI to ensure accuracy.

b. Discuss with general counsel necessary documentation of a review process for AI output.

5 Limiting Use Cases

a. Begin with the firm's existing "acceptable use policy," which deals with access and use of company hardware and software, VPN access, client privacy concerns, etc.

b. Define ordinary use for AI tools, such as writing marketing language vs. extraordinary use such as using AI to produce counsel or professional advice.

c. Ask, who in the firm should have access to AI tools.

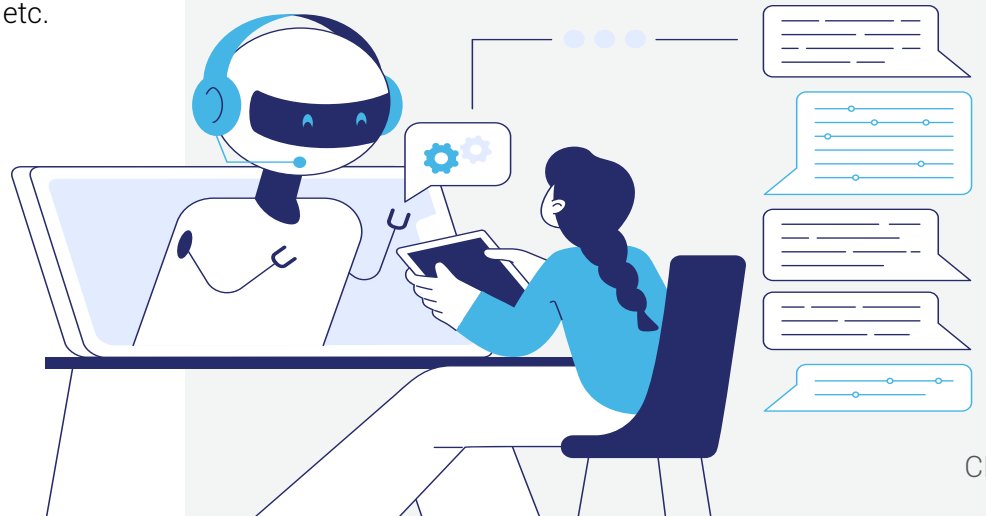
Additional risk considerations, examples and recommendations for mitigation:

Data Exposure and Leaks

The transformative capabilities of generative AI have many risks. One primary concern is the exposure of sensitive information:

Employees Exposing Confidential Information

Case Study: In March 2023, Samsung employees inadvertently exposed confidential data to OpenAI via ChatGPT. This included source code, internal meeting notes, and product roadmaps. ChatGPT uses these inputs to further train their models, which creates the risk that this specific information might leak to the wider ChatGPT user base. Samsung has since banned the use of ChatGPT and other generative AI systems on company-owned devices and internal networks. The leaks have raised concerns about the security risks of using AI tools in the workplace without proper safeguards.



Mitigation: Reviewing the terms of use for the AI systems – some vendors explicitly do not train their models on your data, for others it is part of the T&Cs. Clear policies and procedures in place for the use of AI tools. Regular audits of interactions with external systems, coupled with stringent access controls, to prevent unintentional data sharing.

Data Leakage from LLM Responses

Large Language Models can sometimes disclose sensitive or proprietary information in their outputs. This may result in unauthorized data access, privacy violations, and potential security breaches.

Mitigation: A review layer to vet AI responses, along with the usage of data sanitization protocols and implementing strict user guidelines, can significantly reduce the risk of unintended data disclosure. It's worth noting that some vendors start to indemnify their users from any claims (e.g. Adobe with their visual GenAI).

Insufficient Access Controls

An oversight in the deployment of LLM models can lead to unauthorized users gaining access and model and data exposure.

Mitigation: Multi-factor authentication, role-based access controls, and periodic access reviews ensure only authorized personnel can interact with the AI systems.

Infrastructure Vulnerabilities

The intricacy of enerative AI systems can sometimes introduce specific vulnerabilities that attackers can exploit:

Prompt Injections

Attackers can supply crafty inputs to manipulate LLMs, causing unintended actions or even bypassing filters.

Mitigation: Rigorous input validation and filtering mechanisms. Consider the use of a whitelist approach where only certain predefined types of inputs are permitted.

Inadequate Sandboxing

AI models, when inadequately sandboxed, can allow unauthorized access to the underlying IT environment. Consequences can include:

- Insecure Output Handling
  When an LLM output is accepted without scrutiny, exposing backend systems.
- Unauthorized Code Execution
  Using malformed natural language inputs to execute malicious code or commands.
- Server-side request forgery (SSRF) Vulnerabilities
  Accessing restricted resources, internal services, APIs, or data.
- Exploitable Plugin Design
  Some LLM plugins can have insecure inputs and insufficient access control.

Mitigations:

- Deploy AI models within isolated environments, ensuring they don't have unwarranted access to external systems.
- Regularly update and patch systems to defend against known vulnerabilities.

Improper Error Handling

Errors, when not properly masked, can reveal sensitive data, including personal information.

Mitigation: Customize error messages and debugging output to ensure they remain generic and non-revealing. Conduct thorough testing to identify and rectify information leak points.

Model Denial of Service

Given the resource-intensive nature of generative AI models, attackers might initiate resource-heavy operations causing service degradation or high costs.

Mitigation: Rate-limiting and monitoring systems to detect and alert on unusual activity patterns.

Data Integrity Threats

Generative AI models are only as good as the data they're trained on. Any compromise in data integrity can lead to serious ramifications:

Training Data Poisoning

Deliberate tampering with training data can introduce biases, backdoors, or vulnerabilities.

Mitigation: A secure and controlled environment for data collection and curation. Regularly review and validate datasets to ensure their quality and integrity. Employ anomaly detection to identify unusual patterns in training data.

Model Theft

Theft of proprietary AI models not only results in economic losses but also exposes sensitive information and intellectual property.

Mitigation: Encryption techniques for model storage and transmission. Use secure enclaves or trusted execution environments to protect models during inference. Additionally, utilize model watermarking techniques to trace unauthorized usage.

Compliance and Ethical Concerns

The wide-ranging capabilities of generative AI also introduce potential compliance and ethical challenges:

Over-reliance on Generated Content

Generative AI models can sometimes produce outputs that aren't grounded in their training data, "hallucinations". Excessive trust in LLM-generated content can result in misinformation, miscommunication, legal issues or security vulnerabilities.

Mitigation: Manual review processes for critical AI-generated content. Educate users on the potential risks and encourage a culture of verification.

AI Alignment and Excessive Agency

LLMs can exhibit behaviors misaligned with intended use cases, leading to unintended consequences.

Mitigation: Regularly test and validate AI systems against real-world scenarios to ensure alignment. Restrict excessive permissions and autonomously granted functionalities to LLMs.

Bias and Discrimination Risks

AI responses may sometimes demonstrate biases, risking violations of anti-discrimination laws.

Mitigation: Regularly audit models for bias. Use diverse training datasets and employ fairness-enhancing interventions during model development.

Intellectual Property and Copyright Implications

AI-powered tools are often trained on massive amounts of data and are usually unable to provide sources for their responses. Copyrighted resources, such as books, magazines, and academic journals, may be included in some of the training data. Using AI output based on copyrighted materials without citation could result in legal penalties.

Mitigation: Provide disclaimers when using AI-generated content. Implement fact-check mechanisms to detect potential copyright infringements.

Licensing Restrictions and Content Use

Another consideration involves using content that may be under a license or other agreement. The terms of these agreements might not explicitly permit or could even prohibit the use of such content with AI tools, either fully or without citation.

Mitigation: While this area remains legally ambiguous, it's worth consulting with legal counsel on this matter.

Legalities Surrounding Chatbot Use

Using AI-powered chatbots to answer customer inquiries without proper disclosures may invite legal penalties.

Mitigation: Always disclose AI interactions to users. Stay updated on regional laws and regulations concerning chatbot and AI use.

Open-Source Models Risks

Unvetted Code & Components

Open-source models and their infrastructure are more likely to contain code or components that haven't undergone rigorous security checks. Malicious actors can introduce vulnerabilities or backdoors.

Mitigation: Regularly audit the open-source code for vulnerabilities. Utilize static and dynamic code analysis tools to inspect for potential threats.

Inconsistent Updates & Maintenance

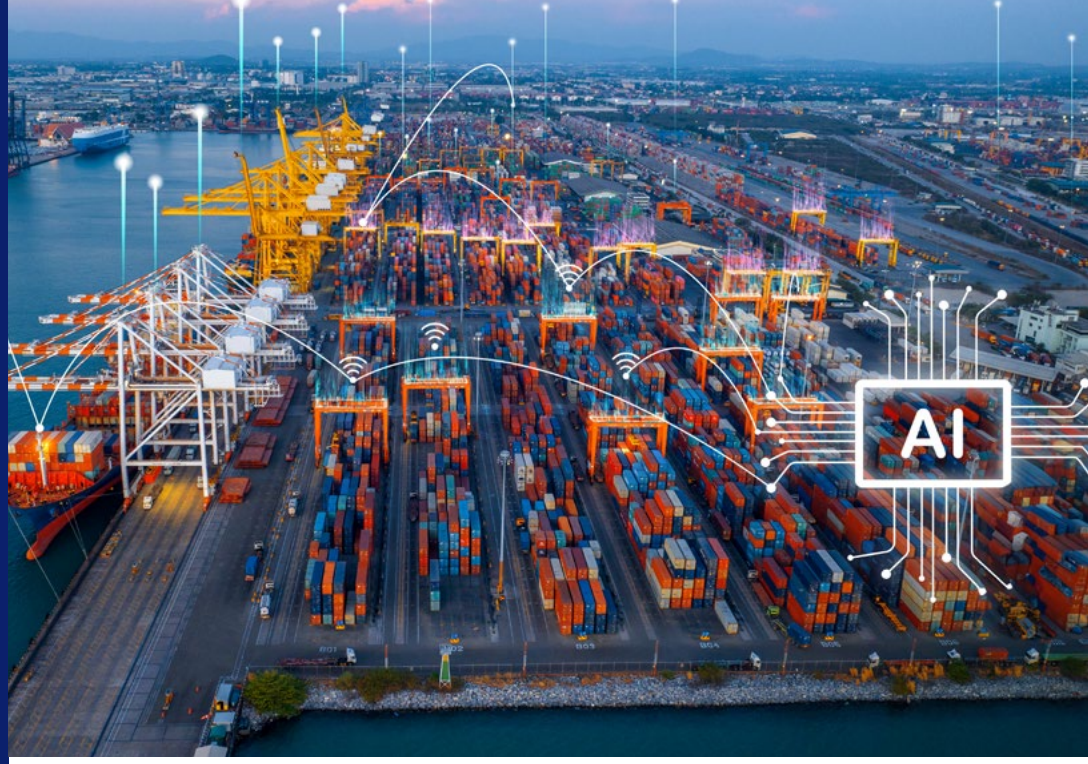
Some open-source projects might not receive regular updates, leaving them exposed to known vulnerabilities.

Mitigation: Ensure you're using actively maintained open-source projects. Regularly check for updates and patches, and integrate them promptly.

Exposure to Pre-trained Data

Open-source models, especially pre-trained ones, might have been exposed to unvetted data sources. This could introduce biases or make the model respond in unpredictable ways.

Mitigation: When possible, fine-tune the open-source models using your own vetted datasets. This can help in overriding potential biases introduced during the initial training.



Additional Attack Vectors and Misuse

The vast potential of generative AI opens avenues for misuse, both in terms of external attacks and malicious applications:

Supply Chain Vulnerabilities

The lifecycle of generative AI can be compromised by vulnerable components or services. The AI ecosystem is rapidly evolving, creating multiple opportunities for malicious actors to attack.

Mitigation: A thorough vetting process for third-party datasets, pre-trained models, and plugins. Ensure a regular update and patching schedule. Conduct security audits on all integrated components.

Data Privacy Concerns

For enterprises who develop their own generative AI models, the process requires large amounts of training data, raising concerns if breaches occur and threat actors gain access to sensitive data, including personal information.

Mitigation: Robust data encryption, both in transit and at rest. Implement strict access controls on data storage locations, and ensure compliance with relevant data protection regulations for the usage, retention, and disposal of stored information.

Malicious Use of Deepfakes

AI's capability to create realistic fake video and voice cloning can be used to impersonate key personnel and obtain sensitive information.

Mitigation: Educate employees about the risks and signs of deepfakes. Implement detection tools that leverage AI to recognize manipulated media.

Personalized Social Engineering Attacks

AI can craft highly targeted phishing or scam messages, enhancing their efficacy.

Mitigation: Train employees and stakeholders on the latest social engineering tactics.

Mitigation Strategies

Beyond addressing specific vulnerabilities, a broader, proactive approach is essential to safeguard against evolving threats in the generative AI landscape:

Continuous Security Training

Educate teams about the evolving threat landscape and best practices in handling and deploying AI tools. A well-informed team can be the first line of defense against potential breaches.

Robust Infrastructure Security

While the focus might be on AI, ensuring that the foundational IT infrastructure is secure is paramount. Regularly update, patch, and monitor underlying systems to deter potential breaches.

Third-party Vendor Audits

Regularly review and audit third-party AI vendors. Ensure they adhere to security best practices and are transparent about their data handling and processing methods. Consider obtaining an assurance report from a licensed CPA firm.

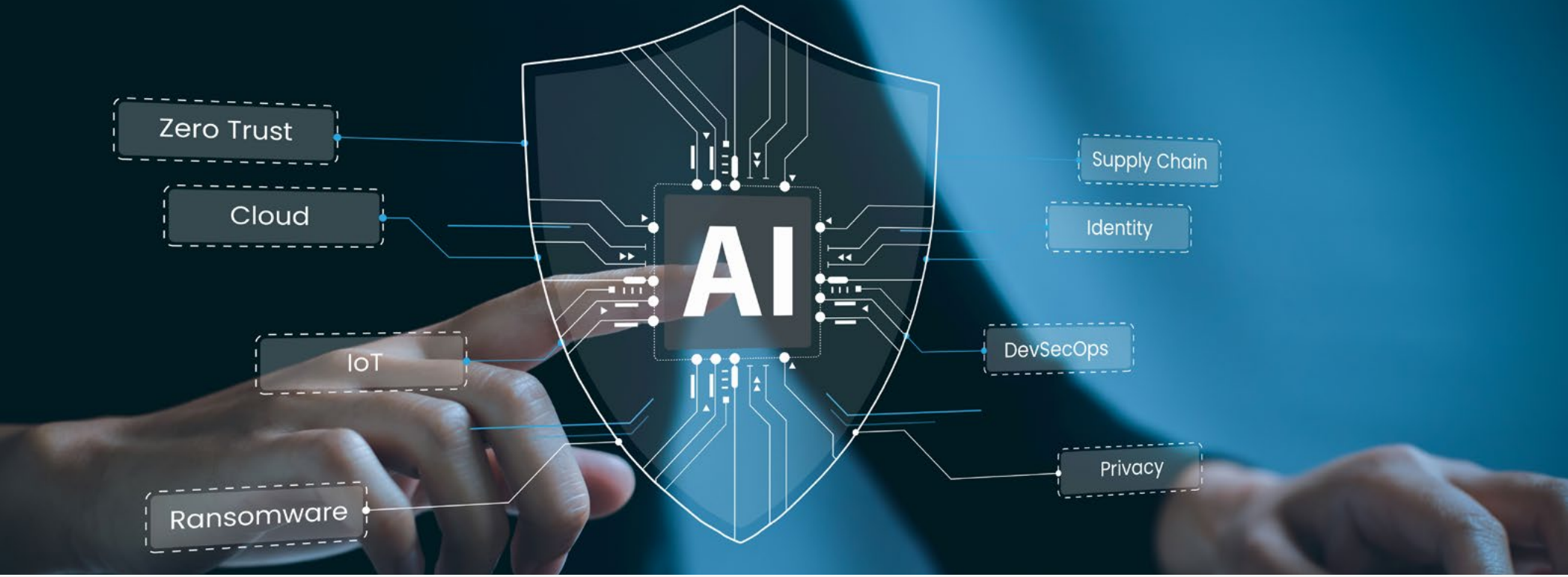
Incident Response Plan

Prepare for potential security incidents by having a response plan in place. This plan should detail steps to contain the threat, communicate with stakeholders, and recover operations.

Feedback Loops

Establish feedback mechanisms with users of AI systems. They can often provide valuable insights into potential system misbehaviors or vulnerabilities.

CPA.com™



Limit AI Autonomy

For high-risk scenarios, consider limiting the autonomy of AI systems. Manual oversight or approval mechanisms can act as safety checks against unintended AI actions.

AI Adoption and Risk Levels

The risks associated with generative AI vary depending on a firm's engagement with the technology. Different threats emerge based on whether a company merely uses tools like ChatGPT for basic tasks, leverages vendor-specific applications for specialized operations, or actively develops and trains their own models using client data.

The Double-Edged Sword of AI

As generative AI tools enhance productivity and open new possibilities, they also present novel risks. It's imperative to approach them with a blend of enthusiasm and caution.

Hacker Productivity

Just as businesses benefit from advancements in AI, so do malicious actors. The tools and methods that enhance legitimate operations can also streamline illicit activities.

Traditional Cybersecurity

Embracing new AI-driven solutions should not come at the cost of sidelining traditional cybersecurity measures. Both old and new defenses must work in tandem to ensure robust protection.