CyberEye

A CISO's Guide to the Assessment of Generative Artificial Intelligence

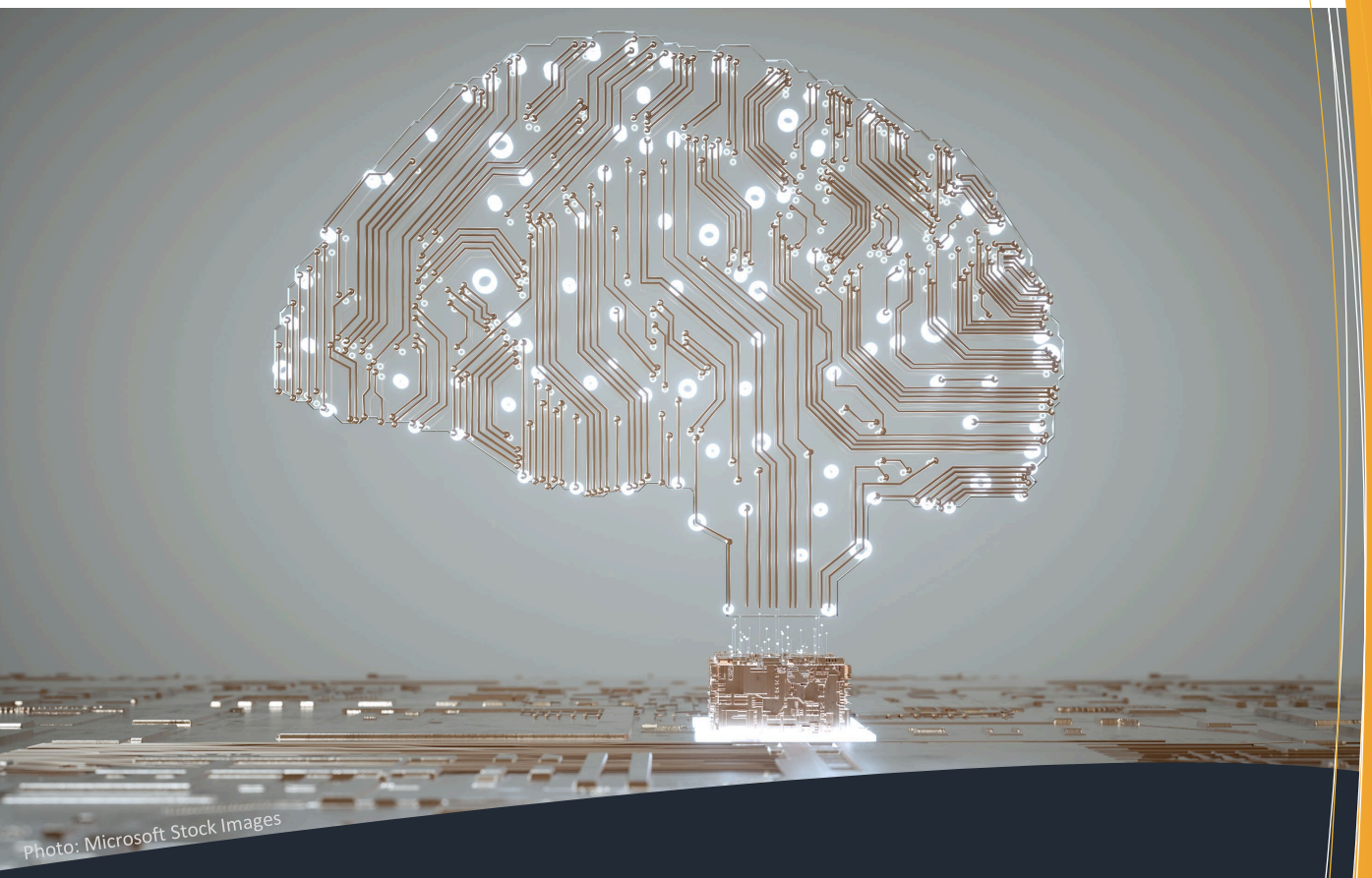


Photo: Microsoft Stock Images

Generative AI

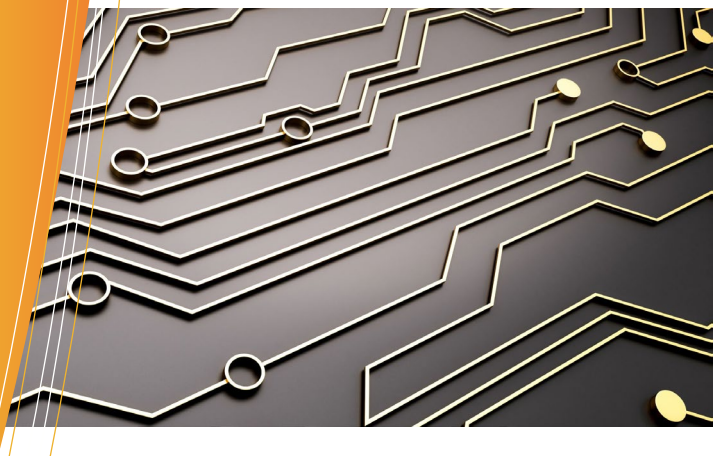
How To Handle Requests For Use in Your Organization

Why Generative AI attracts attention among your staff?

- Novel way to search for information
- Conversational nature of interface
- Automatic email generation
- Ability to generate images from text
- Troubleshooting support for complex business problems

How does the Generative AI service provider benefit?

- Access to massive amounts of user generated data
- Continuous feedback into the AI models
- Ability to leverage user behavior and data for targeted advertising
- Opportunity to sell ongoing subscriptions to users promising enhanced capabilities and integrations



It seems like Artificial Intelligence (AI) is suddenly everywhere. It permeates every aspect of our daily lives from search engines to operating systems and applications, to the local fast-food drive-thru. AI has the potential to profoundly change our society and upend decades of operational practices across industries from healthcare to banking to retail. Many organizations are starting to investigate how Generative AI can benefit them and their workforce. There are several risks that must be weighed before organizations jump into the Generative AI world. Let's discuss Generative AI/Chatbots, their potential benefits, the varied risks, and the marketplace reaction. We'll also look at the recommendations made by the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (RMF).

With ChatGPT and other emerging services, **staff are naturally looking for ways to boost productivity and work quality.**

What is ChatGPT?

ChatGPT and other emerging services like Bard, Sage, Claude, Neeva, Midjourney, Dall-e and Bing Chat are Generative AI services and are typically regarded as AI Chatbots. AI Chatbots rely on large language models (LLM) that are trained on massive amounts of publicly available data that is scraped from the internet as well as other public data sources. These Chatbots offer users opportunities to research, ask questions, have discussions, create graphics, and get advice in a conversational style.

USING THIS GUIDE

This guide provides recommendations for assessing risk and making a business decision as to whether it would be valuable and appropriate to use Generative AI without compromising the safety and security of the business, business data, business partners, valued employees, and end customers (or constituents).

What about the Risks?

The risks of Generative AI to the public and society at large are being fiercely debated, but there is some early consensus in the potential for Generative AI systems to massively upend the labor market. Generative AI technologies have the potential to replace tens of millions of jobs across multiple industries including healthcare, technology, and the creative arts among others[1].

Both public and private sector face unique risks that need to be addressed including safety and reliability, data disclosures, data use, and sale and ownership. Allowing employees access to ChatGPT and other similar technologies runs the risk of inadvertent disclosure of sensitive data as users attempt to solve problems or research issues. These disclosures could be ingested by these systems and then further trained on that ingested data exposing sensitive data to other users and the general public.

What Makes AI Trustworthy?

The National Institute of Standards and Technology[2] considers AI trustworthy if it is:

- **Valid and Reliable**
- **Safe**
- **Secure and Resilient**
- **Accountable & Transparent**
- **Explainable and Interpretable**
- **Privacy-enhanced**
- **Fair with harmful bias managed**

5 Risks of Using Generative AI Every CISO should monitor

RELIABILITY 1
Consumer Beware! Data derived from Generative AI make not be accurate

TRANSPARENCY 2
Look for information on the training data, model structure, intended use cases, and deployment decisions

BIAS 3
Watch out for bias that may call into question the credibility of the data output or create harm

BEHAVIOR TRACKING 4
What the vendor collects from its users and how that will be used is not always clear

EVALUATION 5
Some vendors make it harder than others to really evaluate the service for risk and maintain accountability

Risk No. 1: Reliability

Possible risks specific to Generative AI start with validation. LLM's are often inaccurate and unreliable because of the data they are trained on. The data, not validated for accuracy before being ingested, is likely to spread inaccurate information. When users access these systems, they may assume they are getting accurate information, but too often these services are prone to errors by misstating facts or creating fictional data and presenting them with full confidence. Human reasoning and fact checking are still vastly superior to these chatbots, and without the human element, there is a strong possibility that output from these Generative AI systems could "result in incorrect output that does not accurately reflect real people, places, or facts."[3] The lack of validation without human oversight could lead to concerns over safety of life, health, property, and environment. In some cases, use of Generative AI systems have led to lawsuits from individuals who claim they were defamed by the results these systems have generated.[4]

Risk No. 2: Transparency

Transparency is a risk factor depending on the terms of use for a specific vendor. Does the vendor or organization adhere to open data standards? Are they willing to provide information on the training data, model structure, intended use cases, and deployment decisions? When it comes to a Generative AI systems training data, both the end user and organization should be aware that it may be subject to copyright law and should adhere to all applicable copyright and intellectual property rights laws.

Output may be inaccurate and **unreliable** depending upon the data they are trained on

Risk No. 3: Bias

Another area of concern with Generative AI systems is bias. There is, of course, bias in human decision-making and while bias is not always a negative phenomenon, "AI systems have the potential to increase the speed and scale of biases and amplify harm to individuals, groups, communities, organizations and society."[5] A recent review of Generative Text-to-Image AI systems by Bloomberg found that, systems used by companies like Adobe Inc. and NVIDIA show a large racial and gender bias when given neutral prompts to create images for a dataset of high vs low-paying jobs. "The analysis found that image sets generated for high-paying jobs were dominated by subjects with lighter skin tones, while subjects with darker skin tones were more commonly generated by prompts like fast-food worker and social worker. Categorizing images by gender tells a similar story. Most occupations in the dataset were dominated by men, except for low-paying jobs like housekeeper and cashier."[6]

Risk No. 4: Behavior Tracking

Organizations must consider how their users are being tracked by these systems and how their generated data is being used. OpenAI requires that all users create accounts either through a verified email address or federated sign on using a Google or Apple account. By requiring this, they can store users' conversations and searches, which is turned on by default, and track user behavior at the account level. While it is not clear how all organizations providing Generative AI services are using this data, the future possibilities to leverage this type of user activity is significant and concerning.

Risk No. 5: Evaluation

A key aspect to risk management is the ability for organizations to evaluate systems and make risk-based decisions. Researchers have noted that current Generative AI systems are not easily explainable (how, and why it generates a response) which can inhibit the ability to document systems for end users as well as for the organization to properly audit, measure, and evaluate. There are practical steps that can be enacted to help Generative AI systems become more explainable and therefore accountable. These steps include having a vigorous human review of any generated output, ensuring that the engineers that are building, maintaining, or administering these systems know their algorithms and how they impact the end user. Finally, the data sources used to train these Generative AI systems must be transparent and available for audit.[7]

TERMS OF USE

When it comes to evaluating any of these Generative AI services, it is important to review the Terms of Use to determine if any conditions may be cause for concern. When reviewing the Terms of Use for OpenAI, there are several conditions that may be of concern for organizations considering utilizing ChatGPT or any Generative AI service provided by OpenAI.

OpenAI notes that users are not permitted to represent AI generated output as a human generated response. Any user who uses output generated by their service must provide proper attribution indicating the output was generated from an AI system.

Data storage and Data use by Generative AI companies is an important consideration and in the case of OpenAI, it is clearly stated they reserve the right to use any content created, as a result of a user's input and the system's output, "to provide and maintain the (OpenAI) Services, comply with applicable laws and enforce our policies."[8]

Perhaps most troubling is the language around indemnification and the legal exposure that users are subject to when using services provided by OpenAI. In short, the language indicates users of its service are fully liable for the output that their service provides. In other words, if you as a user create data that emanated from ChatGPT and were then sued because of what you published or used in a work product, it is you as the user who would be responsible for "any claim, losses and expenses (including attorneys' fees) thereof"[9] and not OpenAI who provide the service.

Generative AI Risk vs Traditional Software Risk

Generative AI services present unique risks which are not easily addressed by existing Cyber and privacy frameworks like NIST. Some of these risks include:[10]

- Large language model training data used to build Generative AI systems may not be factual or appropriately representative of the context.
- Changes during model training may alter performance.
- Datasets used to train Generative AI systems may become stale as most do not keep an active connection to data sources.
- Pre-trained models can increase statistical uncertainty.
- Significant privacy risks due to the enhanced data aggregation capability for Generative AI systems.
- Increased opacity from Generative AI systems and services.
- Computational costs for Generative AI systems and their impact on the environment.
- Inability to predict or detect side effects of Generative AI-based systems.

What Approach Should Your Organizations Take?

Many organizations are currently evaluating Generative AI systems for use and considering how this technology will impact them. Some organizations have jumped right into the Generative AI ecosystem while others are taking a wait and see approach. Numerous Fortune 500 companies, including JPMorgan, Samsung, and Apple, have banned employee use of generative AI systems out of concern that it could lead to leaking of corporate secrets and other proprietary information.[11]

CALL TO ACTION

If your organization embraces the use of Generative AI at this early adoption stage, then we recommend you perform due diligence by following these 6 Key Takeaways from the National Institute of Standards and Technology Risk Management Framework for AI and accompanying AI Playbook.

Establish Policy

Document, in policy, the organization's mission and goals for Generative AI technology. Establish a clear set of requirements to guide the organization using a centralized decision authority. Policy can (and should) change over time as regulatory requirements emerge and new risks are revealed. This will enable the organization to establish characteristics of trustworthy Generative AI over time.

Perform Risk Assessments

For each use case it is imperative that a risk assessment be performed and documented to evaluate the data that may be exposed, the business activity impacted, risk to the organization and its stakeholders, including third parties, and their intellectual property rights. These assessments should trigger a response to either mitigate risk or not proceed. Assessment criteria can be derived from the NIST AI Playbook. This may require an update to current risk assessment processes, updated training for staff, and a cultural shift to foster critical thinking, and a safety-first mindset.

Update Procurement Process

Include an evaluation in each procurement of Information Technology (IT) for the use of Generative AI. Initiate a risk assessment and maintain an inventory of Generative AI systems and services for tracking and controlled decommissioning. This includes an assessment of third-party entities and suppliers.

Perform Periodic Reviews of IT Portfolio

Many IT products and services are moving toward Generative AI adoption. While a product or service may not have used Generative AI in the past, that may not be true going forward. Periodically review your IT portfolio for use of Generative AI to proactively address new risks. The value of Generative AI should be revisited periodically as the benefits may degrade over time.

Employ Technical Controls

Ensure that technical enforcement of limitations is in place where a decision has been made to prohibit use of an identified technology. Technical controls may include monitoring user activity or blocking access to the product or service altogether.

Enforce Quality Reviews & Reporting

Establish a mechanism for users to report inaccurate, dangerous, or otherwise untrustworthy output from Generative AI systems to their risk executive so suspect data can be tracked, adjudicated and recorded.

STRATEGIC AND OPERATIONAL LEADERSHIP FOR RISK BASED DECISION MAKING

CyberEye, a commercial services provider in the field of information security Governance, Risk and Compliance, is pleased to present this Guide to Generative AI as a resource for public and private sector organizations evaluating their stance on Generative AI. .

The guide was prepared by our virtual Chief Information Security Officer (vCISO) professionals with an emphasis on AI impact to information security. CyberEye offers access to experienced practitioners who can provide hands-on support and insight on an intermittent, ongoing part-time, or full-time basis, depending upon your needs. Consider using our vCISOs for strategic planning, policy framework development, managing risk, assessments and remediation, and assistance with your incident response planning, testing and execution.

Visit cybereyesolutions.com to learn more.

ABOUT THE AUTHOR

Jospeh B. Isom is an information systems and security professional with over 15 years experience delivering technical solutions that implement sound cyber security practices and empower organizations to meet their business objectives through technology and innovation.

Joseph brings a unique blend of commercial and military experience, relevant certifications, specialization in system and network administration, and foundational cybersecurity experience that offers our clients an operational perspective to their IT and Cybersecurity challenges.

ABOUT CYBEREYE

After over 20 years of specialized experience building transformational cybersecurity programs for the Federal Government and Commercial clients, Kim B. Maurer founded CyberEye, to bring cyber solutions and practical execution plans that will mature Cybersecurity Programs and help organizations make operational risk-based decisions.

CyberEye was founded on the belief that building, implementing and maintaining cyber programs can and should flex with the changes of a growing business. CyberEye offers customizable consulting services for governance, risk and compliance

CyberEye

Refences

[1] Frank, M.R. et al. (2019) 'Toward understanding the impact of artificial intelligence on Labor', Proceedings of the National Academy of Sciences, 116(14), pp. 6531–6539. doi:10.1073/pnas.1900949116.
[2] National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework, (Department of Commerce, Washington, D.C.), NIST AI-100-1. https://doi.org/10.6028/NIST.AI.100-1
[3] OpenAI Terms of use (March 14, 2023) Terms of use. Available at: https://openai.com/policies/terms-of-use (Accessed: 07 June 2023).
[4] Belanger, Ashley. (2023) OpenAI faces defamation suit after ChatGPT completely fabricated another lawsuit. Available at: https://arstechnica.com/tech-policy/2023/06/openai-sued-for-defamation-after-chatgpt-fabricated-yet-another-lawsuit/ (Accessed: 12 June 2023).
[5] National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework, (Department of Commerce, Washington, D.C.), NIST AI-100-1. https://doi.org/10.6028/NIST.AI.100-1
[6] Nicoletti, L. and Bass, D. (2023) Humans are biased. Generative AI is even worse, Bloomberg.com. Available at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/ (Accessed: 09 June 2023).
[7] Satell, G. and Sutton, J. (2019) We need AI that is explainable, auditable, and transparent, Harvard Business Review. Available at: https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent (Accessed: 08 June 2023).
[8] OpenAI Terms of use (March 14, 2023) Terms of use. Available at: https://openai.com/policies/terms-of-use (Accessed: 07 June 2023).
[9] OpenAI Terms of use (March 14, 2023) Terms of use. Available at: https://openai.com/policies/terms-of-use (Accessed: 07 June 2023).
[10] National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework, (Department of Commerce, Washington, D.C.), NIST AI-100-1. https://doi.org/10.6028/NIST.AI.100-1
[11] Ray, S. (2023) Apple joins a growing list of companies cracking down on use of ChatGPT by staffers-here's why, Forbes. Available at: https://www.forbes.com/sites/siladityaray/2023/05/19/apple-joins-a-growing-list-of-companies-cracking-down-on-use-of-chatgpt-by-staffers-heres-why/?sh=46abebea28ff (Accessed: 06 June 2023).

Disclaimer: Unless otherwise noted by a reference, the thoughts, opinions, and ideas presented in this guide are that of the authors and do not represent official guidance from any government entity.