CyberCatch

AI RISK GUIDE

AUTHOR:
SAI HUDA
CHAIRMAN & CEO
CYBERCATCH

*Manage The Risks
For Results*

TABLE OF CONTENTS

© 2023 CYBERCATCH HOLDINGS, INC. ALL RIGHTS RESERVED

BEYOND THE HYPE OF AI

There is a lot of excitement and hype about artificial intelligence (AI) and its potential to transform not only everyday lives but the world of business.

However, AI is a two-sided coin.

On one side there are significant opportunities, while on the other side there are significant risks. Regardless, if one proactively mitigates the risks, then one will be able to take advantage of the opportunities fully and maximize the benefits of AI. First, let's quickly level set and explain in plain language what is AI.

Simply stated, AI is ability of a software to perform tasks that traditionally require human intelligence. Previously, AI was comprised of a model with algorithms that one could train with many data points so it would learn in order to detect patterns and make recommendations and predictions. This was referred to as machine learning and was limited in scope. However, in the fall of 2022, generative AI was given birth with the release of ChatGPT and it completely changed the scope of AI and within two months of its launch, ChatGPT reached 100 million active users.

Generative AI is ability of a software to instead use a deep learning foundation model to train in vast quantities of unstructured, unlabeled data for a wide range of tasks right out of the gate and can be fine-tuned for expanded tasks such as to generate new content. The deep learning foundation model contains an artificial neural network, similar to neurons in the human brain, and can process extremely large and varied sets of structured and unstructured data and can perform multiple tasks all at once. It can classify, edit or summarize data, answer questions, and create new content, images, videos, audio, create computer code, among other tasks.

BEYOND THE HYPE OF AI

Since the original release of ChatGPT by OpenAI, GPT-3.5 and the more advanced GPT-4 was released, which uses Large Language Model (LLM). Anthropic released Claude, which is able to process 100,000 tokens of text, equal to 75,000 words in a minute. Google also released PaLM2, and the AI race has begun.

Since generative AI has the ability to understand natural language, it has the potential to automate work activities and tasks that today take up 60 to 70 per cent of a worker's time, according to McKinsey's research. Thus, it can deliver significantly greater value in use cases involving knowledge workers with higher wages in areas such as customer operations, sales, marketing, risk, compliance, cybersecurity, software engineering and R&D, among others.

According to McKinsey's research, generative AI could add the equivalent of $2.6 trillion to $4.4 trillion annually to the global economy. By comparison, Japan's entire GDP is $5 trillion, Germany's is $3.9 trillion and the United Kingdom's entire GDP is $2.7 trillion.

For example, according to the research, generative AI could add $340 billion - $470 billion in productivity gains in customer operations and $180 billion - $260 billion in risk and compliance business functions across industries.

So, on one side of the coin, AI has the potential to completely transform everyday lives and the world of business, however, on the other side, AI poses significant risks. Specifically, there are five key risks. Let's now explain these five risks.

Generative AI could add upwards of

$4.4 TRILLION

to the global economy.

SOURCE: MCKINSEY RESEARCH

THE FIVE KEY RISKS

There are several risks that AI (generative AI) poses that one must understand and mitigate.

Here's a simple example. Imagine, a staff member at a business uses generative AI to quickly create the content for the web site to promote its products. Everyone is impressed and it generates significant sales. However, a competitor notices the content infringes on its copyrighted content and files a lawsuit. Now the company has to defend itself and prove the content does not violate copyright laws.

This is just a simple example of what could happen without fully understanding the risks and failing to take steps to mitigate. Thus, one can start to sense the possibilities and the mutations with the unintended consequences that lie ahead, no matter the type, size or geographic location of the organization.

A deep analysis of AI models and the various use cases reveals there are five key risks that one must fully understand and proactively mitigate.

Here they are listed, followed by an explanation in plain language and in practical terms for executive decision making:

- SHADOW
- SECURITY
- INACCURACY
- BIAS
- HALLUCINATION

It should be noted that security is the most significant AI risk and requires focused risk mitigation.

The playbook provided after the explanation of the five risks provides a "how-to" roadmap to manage security risk.

RISK

SHADOW

RISK

SHADOW

HOW IT HAPPENS

Shadow risk is the risk that employees in an organization use AI without approval or use it in an unapproved manner, adversely impacting the organization. The adverse impact can range from inadvertent leakage of company intellectual property to violations of data privacy and copyright laws to creating security holes.

It is a risk similar to “shadow IT “ risk where employees use unapproved hardware or software unknown to the organization, causing a variety of risks, including security risk.

For example, Samsung after discovering that the use of generative AI by employees led to accidental leak of internal sensitive data and intellectual property, decided to ban it until proper measures were put in place to safely use AI. It found out that it is difficult to retrieve or delete the transmitted data on external servers and it could be disclosed to other users.

Apple, JPMorgan, Verizon, Walmart are some of the other larger organizations that have forbid staff from unauthorized use of generative AI tools such as ChatGPT. However, majority of organizations worldwide have yet to recognize shadow risk.

Shadow risk can happen to any organization, no matter the size, and have an adverse impact.

Some larger organizations are beginning to recognize shadow risk and have started to take action.

RISK

SECURITY

RISK

SECURITY

HOW IT HAPPENS

Security risk is the risk that cyber attackers break into the network and access the AI model and related database and content, exfiltrate a copy, install ransomware to encrypt and lock it down and or manipulate the model to adversely impact its use or the results, to harm the organization.

Security risk is inherently increased with use of AI since the AI model and related database and content represents a valuable, strategic IT asset and expands an organization's attack surface.

Depending on the scope and extent of an organization's use of AI, the cyber attackers could range from a hostile nation state directly or a sponsored threat actor group, interested in stealing the valuable intellectual property created from the AI model of national security importance. Also, criminal gangs, ranging from some of the current known ransomware gangs to new gangs with an AI focus. These gangs would be interested in making money from the cyberattack by stealing a copy of the AI model, related database and content and installing ransomware and extorting a ransom payment or selling the valuable intellectual property in the dark web for financial gain.

The cyber attacker could also be a threat actor group that uses a prompt injection to manipulate the AI model with new instructions and trick it to generate content unintended by the organization, ranging from privacy violations consisting of disclosure of personally identifiable information to misinformation, disinformation, deepfakes or hate speech, that could cause significant harm.

Security risk is the most significant of the five AI risks posed.

Organizations should also recognize that cyber attackers are going to use AI for their benefit to enhance their attack tactics, techniques and procedures (TTPs) to attain greater success.

For example, generative AI can be used by cyber attackers to efficiently and precisely:

- Scan Internet-facing IT assets and identify specific technologies, platforms and specific attack surface vulnerabilities.
- Create sophisticated and polymorphic (detection evading altering) codes to exploit identified vulnerabilities effectively and with speed.
- Create spear phishing emails with flawless language, context and personalization to deceive the recipient.

Hence, AI risk, with its inherent security risk, must be battled with equal force by using AI for cybersecurity use case.

Cyber risk mitigation is an optimal use case for using AI to combat the increased level of risk posed.

Security risk is

INHERENTLY INCREASED

with the use of AI.

RISK

INACCURACY

RISK INACCURACY

HOW IT HAPPENS

Inaccuracy risk is the risk that the AI model will produce incorrect answers or inaccurate results or results that cannot be supported with evidence or facts.

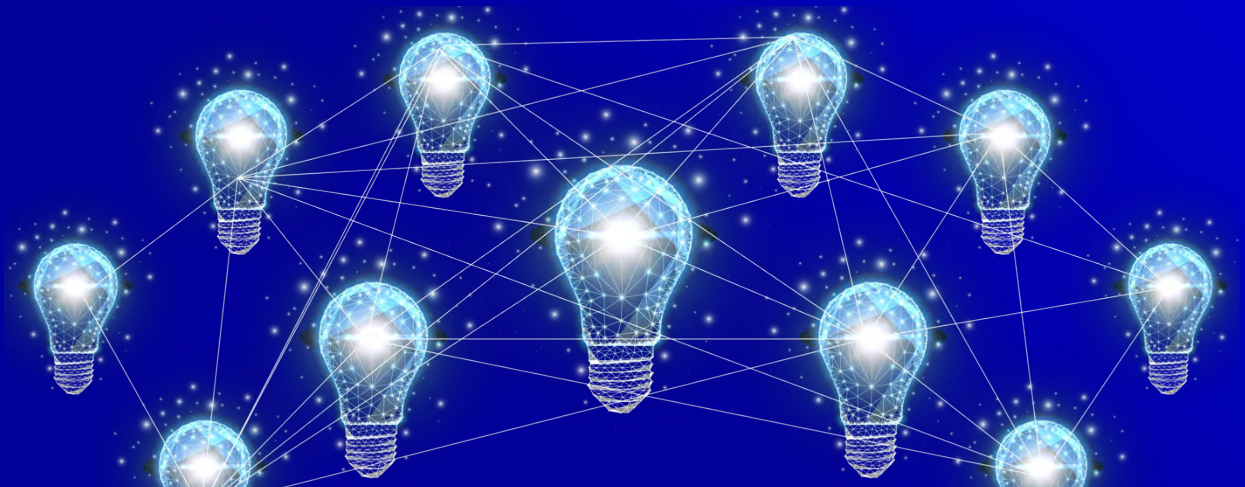
The model could possibly produce different answers to same prompts or questions, creating concerns about the reliability of the model and the output.

The AI model could also produce content that may infringe on another organization or person's intellectual property rights and or violate copyright laws.

It should be noted that generative AI uses a deep learning foundation model to train in vast quantities of unstructured, unlabeled data for a wide range of tasks and can be fine-tuned for expanded tasks.

Additionally, the AI model contains an artificial neural network, similar to neurons in the human brain, and can process extremely large and varied sets of structured and unstructured data and can perform multiple tasks all at once.

Regardless, it is a model after all, and there is a risk that it may produce inaccurate results or results that violate the law, without some level of human quality assurance review and checks and balances.



RISK

BIAS

RISK

BIAS

HOW IT HAPPENS

Bias risk is the risk that the AI model will generate unintended algorithmic bias in the content or results produced, from unintended training data fed into the model or from improper model design, causing harm to users or consumers.

For example, let's take a customer operations center that uses AI model to make supervisory or managerial decisions to provide customers a discount or rebate when a customer complaint is escalated. Unless the model is designed properly, fed appropriate training data and validated against bias, it could end up favoring those without an accent and command of the language, versus those with an accent or English as a second language, resulting in discrimination based on race or ethnicity.

In another example, let's take the use of AI to generate text-to-image ads. Unless the AI model is fed appropriate training data and validated against bias, it could generate stereotypical images in the ads, potentially causing unfair treatment by discouraging certain consumers or users.

Recently, the Federal Trade Commission (FTC) in the U.S. issued a warning regarding unfair business practices that may be created unintentionally from use of AI models without adequate human oversight and checks and balances.

Unless the AI model is fed appropriate training data and validated against bias, it could generate unintended algorithmic bias, causing harm to consumers or users.

RISK

HALLUCINATION

RISK

HALLUCINATION

HOW IT HAPPENS

Hallucination risk is the risk that the AI model will not understand the prompts or questions or will misinterpret and generate content or results that are invented, are not real or are incoherent or are nonsensical.

The AI model learns by analyzing massive amount of digital text from the Internet and since the Internet is filled with untruthful information, it could repeat the same untruths or make things up.

So, it is possible the AI algorithm and deep learning neural network will produce output that are not real and are not based on any training data fed into the model.

For example, a user asked ChatGPT how James Joyce and Vladimir Lenin first met, which never happened, the AI model replied, “James Joyce and Vladimir Lenin met in Zurich, Switzerland in 1916.”

While OpenAI spent six months to improve GPT-4 and it is 40% more likely to produce factual responses than GPT-3.5, a high level of risk remains.

MITIGATING AI RISK

AI is a two-sided coin, with one side containing significant value generating opportunities, while on the other side there are significant risks.

However, the key to success is to proactively mitigate the risks so the organization can take advantage of the opportunities.

Following is a playbook to mitigate the AI risks:

1. At the board and executive level, decide whether the time is right to use AI in the organization and if so for what specific use cases. AI is not a panacea and appropriate for all use cases. However, there are certainly value creating opportunities and relevant use cases for an organization to benefit from the use of AI.
2. Create a policy to ban shadow AI and disseminate to all employees, explaining shadow risk.
3. Create a policy and procedure for responsible use of AI in the organization with scope, permissible use cases, roles, responsibilities, process, and risk management, covering the five AI risks, controls, testing, and reporting outlined. Disseminate to all employees.
4. Implement a cybersecurity program over AI, implementing the controls in NIST Cybersecurity Framework (CSF) to cover the scope of the use of AI in the organization.
5. Implement an ongoing AI compliance testing program to ensure AI risks are mitigated and remain mitigated, and report periodically to executive management and the board for proper risk management oversight.

As explained earlier, security risk is the most significant of the five AI risks posed and requires focused risk mitigation.

Here is a playbook to mitigate security risk:

1. Think like the cyber attacker and develop an enterprise-wide attack surface map and identify the AI model, related database and IT assets, in the map. Cover third-party suppliers in the scope.
2. Identify TTPs a cyber attacker could use to breach the AI model, related database and IT assets, and the cybersecurity controls in place to prevent, detect and respond. Benchmark to NIST CSF to identify non-compliance, security holes and blind spots. Remediate promptly to eliminate security holes.
3. Perform a red team pen-test to validate security risk is mitigated over the AI model, related database and IT assets. Remediate promptly any deficiencies. Implement ongoing automated or manual testing of cybersecurity controls to prevent recurring security holes that can be exploited by cyber attackers.
4. Provide training to all employees involved with AI model on the AI risks, and in particular security risk, so they can serve as a strong human firewall and assist with continuous risk mitigation.
5. Implement a cyber incident response plan covering AI model, related database and IT assets or ensure current cyber incident response plan's scope covers AI. Test the efficacy of the plan, simulating an attack targeting the AI model, to identify security holes and blind spots and remediate promptly.

CyberCatch

HOW CYBERCATCH IS USING AI TO TRANSFORM CYBER RISK MITIGATION

CyberCatch is transforming cybersecurity with its patented, proprietary, artificial intelligence-enabled (AI) Software-as-a-Service (SaaS) solution that enables continuous compliance and cyber risk mitigation for organizations in critical segments, so they can be safe from cyber threats.

The CyberCatch platform solution is unique and industry-leading because it focuses on solving the root cause of why cyberattacks are successful: security holes from control deficiencies.

The CyberCatch platform solution first helps implement all mandated and necessary controls with an AI-enabled automated cybersecurity advisor, guiding the organization to attain compliance in two weeks, instead of industry-average three months or longer.

Then the platform automatically and continuously tests the controls from three dimensions (outside-in, inside-out and social engineering) to find control failures so one can fix them promptly to stay compliant and safe from attackers. An automated pen-tester performs 36 different pen-tests mapped to MITRE ATT&CK TTPs, simulating a cyber attacker to detect security holes for prompt remediation.

The AI-enabled automated cybersecurity advisor continuously guides the organization for continuous cyber risk mitigation.

Additionally, an AI-enabled security coach guides the organization's employees as part of ongoing virtual reality security awareness learning game, and tests susceptibility to social engineering to strengthen the human firewall and make sure it is not the weakest link.

CyberCatch also provides industry-leading expert security advisory services to its customers ranging from AI risk management advisory to mitigate risks from use of AI in an organization, cybersecurity maturity assessments, cyber incident response advisory, red team pen-test to fractional CISO services.

CyberCatch will continue its leadership in the industry by continuing to maximize use of AI for the benefit of its customers.

CyberCatch is transforming cybersecurity with its patented, proprietary, AI-enabled SaaS solution that enables continuous compliance and cyber risk mitigation for organizations in critical segments, so they can be safe from cyber threats.

ABOUT THE AUTHOR

Sai Huda

Founder, Chairman &
CEO of CyberCatch

Sai Huda is the founder, chairman and CEO of CyberCatch. He is a globally recognized risk management and cybersecurity expert and frequent keynote speaker at industry conferences. He is co-author of Canada's national cybersecurity standard and author of the best-selling book, Next Level Cybersecurity and the Cyber Risk Guide for Board and Executive Management. He is former founder, chairman and CEO of Compliance Coach, which was acquired by FIS (NYSE: FIS), a FORTUNE 500 company. He is former GM, Risk, Information Security and Compliance, FIS, and under his leadership FIS attained the number one ranking in RiskTech 100.

ABOUT CYBERCATCH

CyberCatch is a cybersecurity company that provides a proprietary, AI-enabled Software-as-a-Service (SaaS) solution that enables continuous compliance and cyber risk mitigation to organizations in critical segments, so they can be safe from cyber threats.

The CyberCatch platform focuses on solving the root cause of why cyberattacks are successful: security holes from control deficiencies.

It first helps implement all mandated and necessary controls, then the platform automatically and continuously tests the controls from three dimensions (outside-in, inside-out and social engineering) to find control failures so one can fix them promptly to stay compliant and safe from attackers.

LEARN MORE: www.cybercatch.com

CyberCatch

AI RISK GUIDE

WWW.CYBERCATCH.COM