Deloitte.

Safeguarding Generative Artificial Intelligence (AI) with cybersecurity measures

Risk insights and building blocks for secure Generative AI solutions

September 2023

Table of contents

Introduction

2022 was a watershed in the history of Artificial Intelligence (AI). While the journey started in 1932, it has gradually become all-pervasive − first with digital assistants from various multinational technology conglomerates and now with the release of Generative AI. This phase is incredible because it affects most businesses and personal interactions. While the world celebrates the coming of age of Generative AI, one needs to make certain considerations to ensure that the scale and impact are progressive for individuals, organisations, and society alike.

Generative AI has many positive implications and could have the ability to transform the way we do business. Some of the most important aspects include the following:

Intelligent Information Technology (IT) – Transforming how IT is structured, how software development is done, and how IT is enhanced and supported.

Intelligent products – Enhancing sensor-infused products using Generative AI, which can have huge implications across several industries.

Intelligent operations – Remodeling operations with a greater emphasis on Generative AI-derived inputs that could help make operations nimble and agile.

As a result, it will lead to the following outcomes:

1. Improving productivity
2. Increased customer satisfactions
3. Propelling Research and Development (R&D)
4. Creating new revenue streams and business models

While organisations and businesses adopt the Generative AI, Cybersecurity is paramount. Necessary controls should be implemented to ensure that investments deliver the right business results to organisations while maintaining individuals' privacy and confidentiality. Additionally, a lack of adoption of the Security by Design (SbD), Privacy by Design (PbD), and Ethical by Design (EbD) concepts could lead to exposure and risks to data being used and training of models adopted. Finally, security technology needs to keep pace with the development of Generative AI.

This point of view (POV) provides insights into certain cybersecurity considerations for Generative AI and the necessary controls organisations should consider while building these systems.

Evolution of Generative AI

Most of the time, Generative AI is considered relatively new. Contrary to the belief, Generative AI is deep-rooted in history and innovation.

1932

Georges Artsrouni invented a machine reportedly called the "mechanical brain" to translate languages on a mechanical computer encoded onto punch cards.[1]

1973

Harold Cohen, a painter and professor, collaborated with a programme called AARON to produce art autonomously. Paintings are done in Cohen's style.

1989

Yann Lecun, Yoshua Bengio, and Patrick Haffner demonstrated how Convolutional Neural Networks (CNNs) can recognise images.

2000

Researchers from the University of Montreal published "a neural probabilistic language model", which suggests a method to model language using feed-forward neural networks.

2006

Data scientist Fei- Fei Li set up the ImageNet database that laid the foundation for visual object recognition.

2011

Apple released *Siri,* a voice-powered personal assistant that can generate responses and take actions in response to voice requests.

2014

Ian J. Goodfellow and colleagues published the first paper on Generative Adversarial Networks (GANs) that can determine if an image is real or fake.

2017

Google researchers developed the concept of transformers in the seminal paper "Attention is all you need." The paper inspired subsequent research into tools that could automatically parse unlabeled text into Large Language Models (LLMs).

2018

Google researchers implemented transformers into BERT, trained on over 3.3 billion words. It can automatically learn the relationship between words in sentences, paragraphs, and even books, and predict the meaning of text. Google DeepMind researchers developed AlphaFold to predict protein structures that laid the foundation of Generative AI.

2022

Open AI released ChatGPT in November to provide a chat-based interface to its GPT 3.5 LLM. It attracted more than 100 million users in two months, representing the fastest-ever consumer adoption of a service.

2023

Adobe launched Firefly, a family of Generative AI models tailor-made for creative professionals, with built-in guardrails for safety and copyright standards.

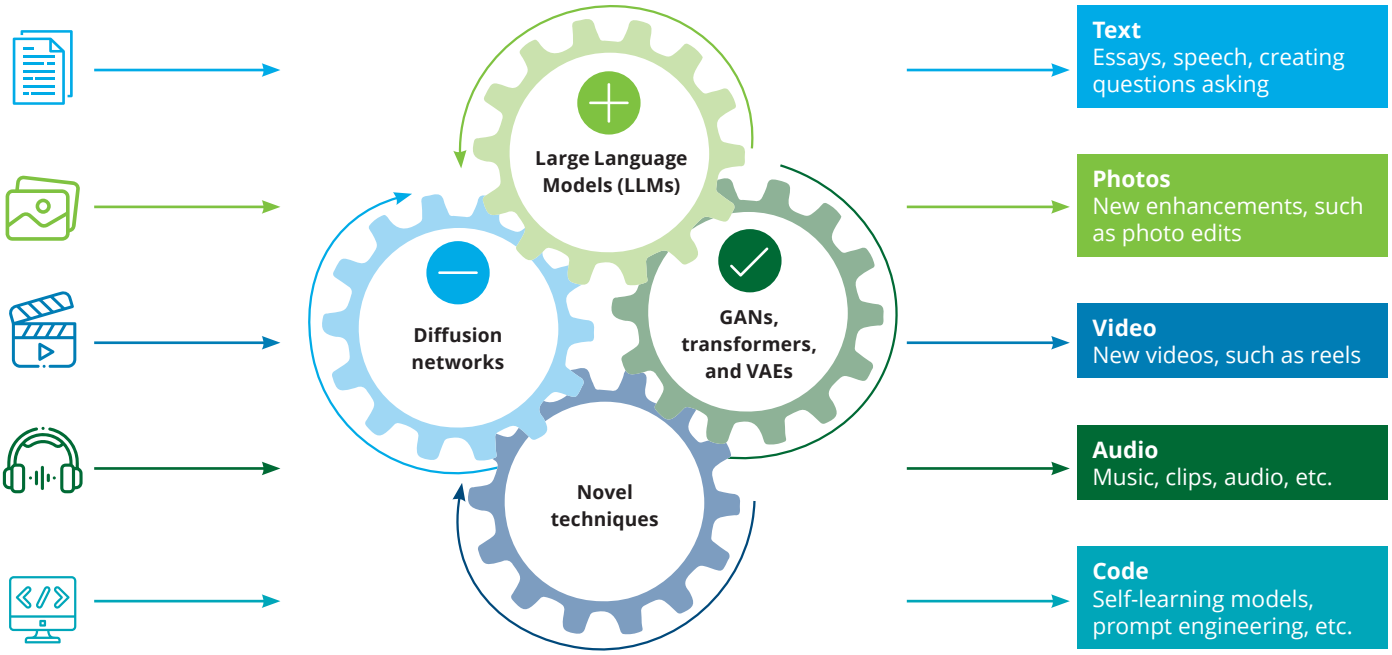
Google released Bard, a Generative AI chatbot built on 137 billion parameters and embeds Generative AI capabilities into its workshop products.

[1] What is Generative AI? Everything you need to know

What constitutes Generative AI

Generative AI learns from various inputs to generate its outputs.[2]

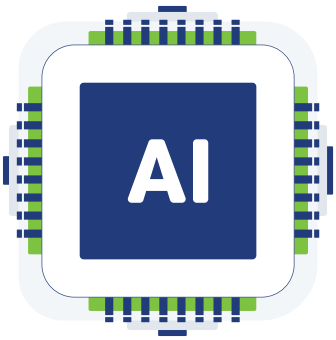
Data can vary from text and photos to videos, audio, and codes. Using these wide varieties of datasets, Generative AI creates novel outputs. Generative AI models use technological elements, such as LLMs, diffusion networks, GANs, transformers and Variational Auto Encoder (VAEs), and other novel techniques to identify patterns and structures within existing data to generate new content.



[2] Generative AI – What is it and How Does it Work? (nvidia.com)

How does Generative AI work?

A simplified view



AI infrastructure

Computer power

Powering our journey to tomorrow

The scale of compute capacity required to train and process foundation models necessitates the use of leading GPUs (e.g., A100 NVIDIA) and TPUs (e.g., Google TPU v4) on scalable infrastructure.

Cloud and data platforms

Data

Training on the world's knowledge

Foundation models are trained on petabytes worth of global data to shape understanding, tone, and behaviour while considering human communication styles.

Generative AI models

Prediction

Why do these applications seem so human?

Similar to traditional AI, foundation models predict outputs based on inferences on the inputs it receives. However, through fine-tuning, prompt engineering, and adversarial training, these models generate outputs based on their understanding of human communication.

Models

What are foundation models?

OpenAI's GPT-4 and NVIDIA's Megatron are two examples of foundation models, specifically Large Language Models (LLMs) that use deep learning to process huge amounts of data. These models form "memories" on the input datasets through tokenisation, thereby shaping models' parameters. Common foundation model architectures, such as transformer and diffusion, drive modalities for each model.

Generative AI applications

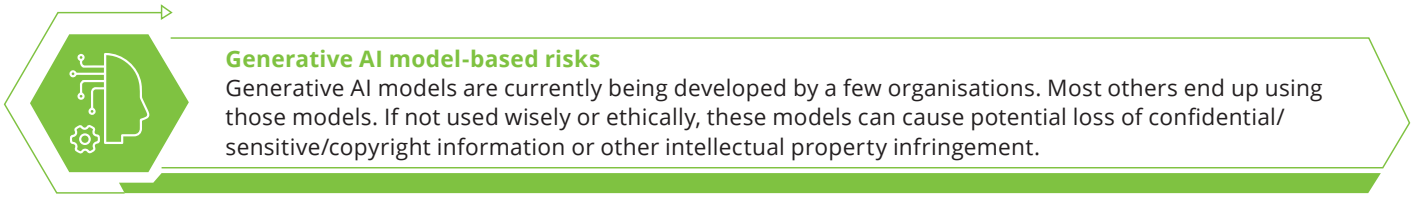
Output

Applications – What we see

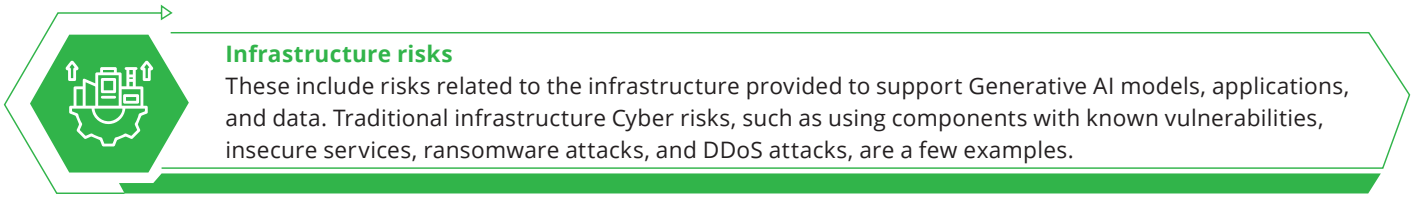
Generative AI applications generate content across various modalities (e.g., text, image, video, and audio).

Six categories of Cyber risks with Generative AI

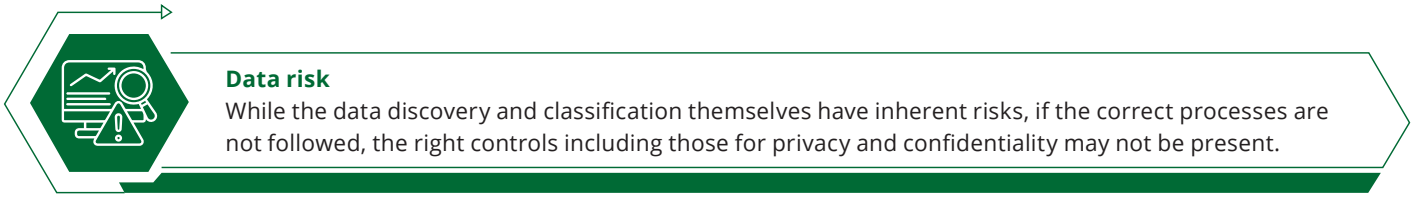
While there are many Cyber risks related to Generative AI, we have tried to group them into six categories.



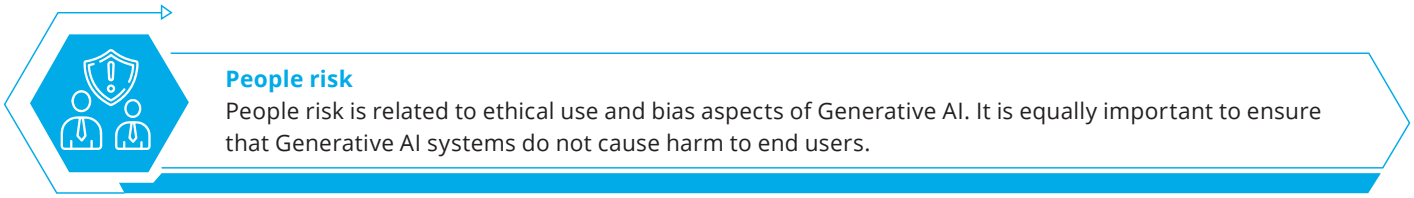
Generative AI model-based risks
Generative AI models are currently being developed by a few organisations. Most others end up using those models. If not used wisely or ethically, these models can cause potential loss of confidential/ sensitive/copyright information or other intellectual property infringement.



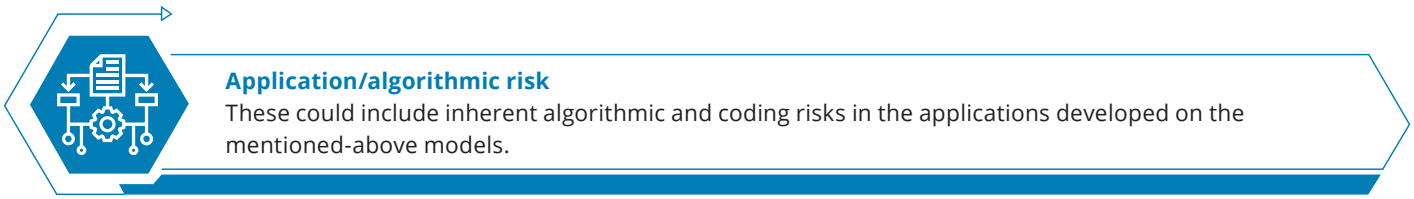
Infrastructure risks
These include risks related to the infrastructure provided to support Generative AI models, applications, and data. Traditional infrastructure Cyber risks, such as using components with known vulnerabilities, insecure services, ransomware attacks, and DDoS attacks, are a few examples.



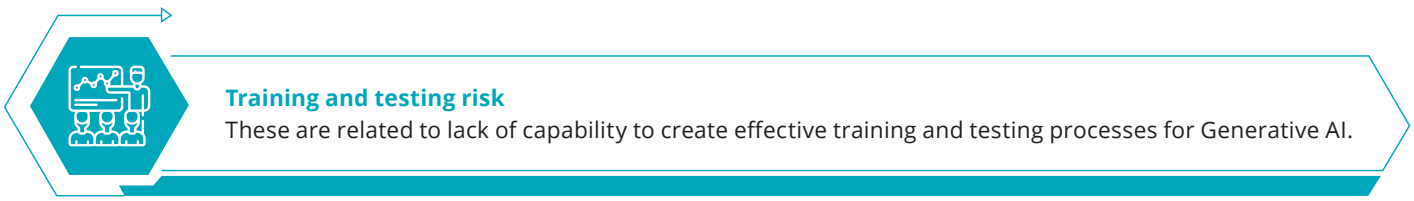
Data risk
While the data discovery and classification themselves have inherent risks, if the correct processes are not followed, the right controls including those for privacy and confidentiality may not be present.



People risk
People risk is related to ethical use and bias aspects of Generative AI. It is equally important to ensure that Generative AI systems do not cause harm to end users.



Application/algorithmic risk
These could include inherent algorithmic and coding risks in the applications developed on the mentioned-above models.



Training and testing risk
These are related to lack of capability to create effective training and testing processes for Generative AI.

Illustrative Cyber risks of Generative AI

While there may be myriad risks from a cyber perspective, some key risks you should be aware of while using Generative AI are provided below.

Membership inference

Inferring the presence of specific data points in the training set by querying the AI model and compromising data privacy.

Insecure output Handling

This vulnerability occurs when a Generative AI output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences, such as XSS, CSRF, SSRF, privilege escalation, or remote code execution.

Model denial of service

Attackers cause resource-heavy operations on Generative AI, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and the unpredictability of user inputs.

Supply chain vulnerabilities

Generative AI application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plug-ins can increase vulnerabilities.

Sensitive information disclosure

Generative AI may inadvertently reveal confidential data in its responses, leading to unauthorised data access, privacy violations, and security breaches. Implementing data sanitisation and strict user policies to mitigate these risks is crucial.

Insecure plug-in design

Generative AI plug-ins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences such as remote code execution.

Prompt injection

Prompt Injection manipulates a Generative AI through crafty inputs, causing unintended actions by the Large Language Models (LLMs). Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

Excessive agency

Generative AI-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to LLMs-based systems.

Overreliance on Generated AI

Systems or people overly depending on Generative AI without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by Generative AI.

Model theft

Model Theft could involve unauthorised access, copying, or exfiltration of proprietary Generative AI models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

Training data poisoning

This occurs when Generative AI training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behaviour. Sources include Common Crawl, WebText, OpenWebText, and books.[3]

Deepfakes

Deepfake technology[4] has advanced to the point where it can be used in real-time, enabling fraudsters to replicate someone's voice, image, and movements in a call or virtual meeting.

[3] https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0.pdf

[4] https://www.latimes.com/business/technology/story/2023-05-11/realtime-ai-deepfakes-how-to-protect-yourself#:~:text=Cybersecurity%20experts%20say%20deepfake%20technology,a%20call%20or%20virtual%20meeting.

Industry-wise use cases of Generative AI, Cyber Risks, and Controls

Generative AI's potential extends to almost every industry, as it provides a wide range of automation and enhances creative and data-driven processes. However, with its various use cases, it also introduces Cyber risks.

A few use cases of Generative AI across key industries, along with the possible Cyber risks and mitigation steps, are provided below:

** This is not an exhaustive list.

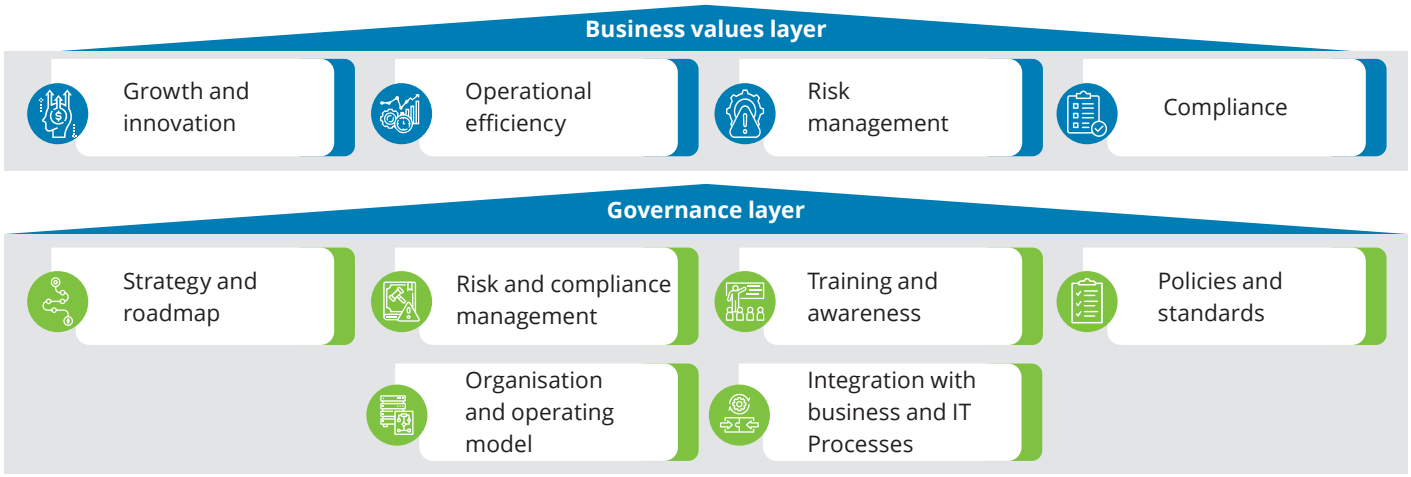
Industry	Use cases of Generative AI	Description of use case	Cyber risk	Mitigation/controls
Consumer	Blog and social media content writing	**Content generation:** Generative AI can be used to generate a variety of content types, including text, photos, and videos. It can, for example, be used to generate personalised product descriptions, blog entries, or even fictional stories.	**Misinformation and fake content:** Generative AI can produce false information, such as fake news articles, photos, and videos. This raises the possibility of distributing inaccurate information, altering public opinion, or carrying out social engineering attacks.	**Fact-checking and verification:** Establish partnerships with reputable fact-checking organisations to verify the accuracy of the content generated. Develop automated systems to cross-reference the generated information with reliable sources, to help identify and flag potential misinformation.
Government and Public Services	Social services and welfare	Generative AI can assist in personalised service delivery in healthcare, social welfare, or education. It can analyse individual data to recommend suitable programmes, interventions, or support services based on specific needs.	**Bias and discrimination:** Generative AI models trained on biased or incomplete data may inadvertently perpetuate biases or discrimination in social service delivery. This could lead to unfair or discriminatory outcomes, disadvantaging certain individuals or groups based on their demographic or socio-economic characteristics.	**Transparent decision-making:** Enhance the transparency of Generative AI systems by explaining decision-making processes. Use techniques such as Explainable AI (XAI) to make the generated outcomes more understandable to citizens, fostering trust and accountability.
Energy, Resources, and Industrials	Energy demand forecasting	Generative AI models can analyse historical energy consumption data, weather patterns, economic indicators, and other relevant factors, to help forecast future energy demand. Accurate demand forecasting helps utilities and energy providers optimise resource allocation, plan for peak demand periods, and enhance energy distribution.[5]	**Manipulation of forecasting data:** Adversaries may attempt to manipulate or tamper with the data used for energy demand forecasting. By injecting false or misleading information into the dataset, they could manipulate the forecast demand, potentially leading to inefficient resource allocation, financial loss, or disruptions in energy supply.	Implement robust data security measures, including encryption, access controls, and secure storage to protect the confidentiality and integrity of the data used for forecasting. Use anomaly detection and outlier analysis techniques to identify and mitigate potential data manipulation attempts.

[5] How generative AI can boost productivity in enterprises and industries

Industry	Use cases of Generative AI	Description of use case	Cyber risk	Mitigation/controls
Financial Services	Financial forecasting	By learning from historical financial data, Generative AI models can capture complex patterns and relationships in the data, enabling them to make predictive analytics about future trends, asset prices, and economic indicators.[6]	**Model poisoning:** Threat actors may manipulate the training process of Generative AI models by injecting malicious data or disturbing the training data to undermine the accuracy of forecasts.	Regular model auditing and monitoring of adversarial activities are a few mitigating controls to help combat the "poisoning of the data".
Technology, Media, and Telecommunications	User interface design	Generative AI can help in User Interface (UI) design by providing automated suggestions for layouts, colour schemes, and component placement based on user requirements or predefined templates. This can help developers in rapid prototyping.[7]	**SSRF vulnerabilities:** They allow the exploitation of Generative AI models by performing unintended requests or accessing restricted resources, such as Application Programming Interfaces (APIs) or internal services that may lead to wrong designs.	Rigorous input validation and regular audit network/ application security.
Life Sciences and Health Care sector	Drug discovery	Generative AI can be used to streamline drug discovery and development by identifying potential drug candidates and testing their effectiveness before moving them for other trials.[8]	**Intellectual property theft:** Generative AI models are trained on extensive datasets, which may include proprietary or patented information. Unauthorised access to these models or their outputs could result in intellectual property theft, where competitors could access confidential drug discovery processes, formulas, or compounds.	Intellectual risks in Generative AI can be mitigated using multiple strategies together, such as encryption, secure data hosting, access controls, water-marking, digital signatures, or content fingerprinting.

[6] Generative AI in the Banking and Finance Industry
[7] Generative AI: The Next Frontier in Telecom Innovation
[8] Generative AI Healthcare Industry: Benefits, Challenges, Potentials

Building blocks for secure Generative AI solutions



Five pillars of responsible Generative AI adoption for a secure AI ecosystem

Policy and principles
Develop policies and principles focusing on transparency, accountability, and safety for developing/ buying Generative AI technologies.

Governance
Establish a governance programme to manage Generative AI technology.

Risk assessment
Conduct risk assessments to identify potentials risks, safety and bias issues.

Update your policy regularly, with mitigation controls.

Continuous evaluation and monitoring
Regularly assess fairness, bias, accuracy, and impact on users.

Infrastructure monitoring and security
Set up processes for infrastructure monitoring, to help address vulnerabilities and incorporate the most recent developments in Generative AI security.

Awareness
Encourage the responsible use and discourage malicious or unethical applications of Generative AI.

Planning
Create a scalability plan that outlines the steps, benchmarks, and risk mitigation measures to scale Generative AI systems.

Model robustness
Verify the robustness and generalisation capabilities of Generative AI models during scaling.

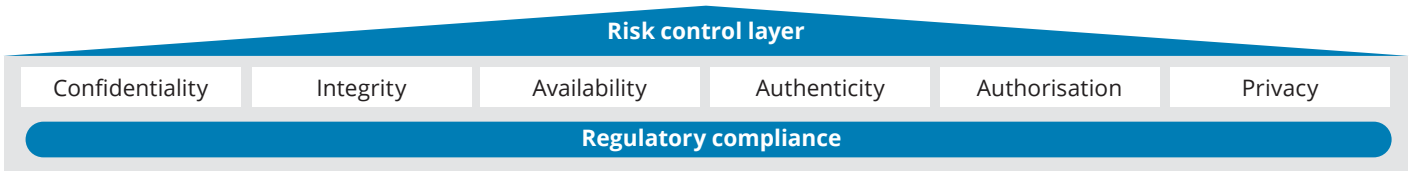
Risk assessment
Conduct a risk assessment to identify potential risks, such as data breaches, or unintended consequences, as a result of scaling.

Impact assessment
Conduct impact assessments to evaluate the potential impacts, risks, and considerations associated with customising Generative AI systems.

Transparency
Maintain transparency while customising Generative AI systems, to enable users to trust the customised system.

Decommissioning policies
Create clear protocols and procedures for shutting down or decommissioning Generative AI systems. Ensure to address data retention issues and privacy problems.

Safeguard against the malicious use of Generative AI technologies, such as models, data, and associated infrastructure after termination.



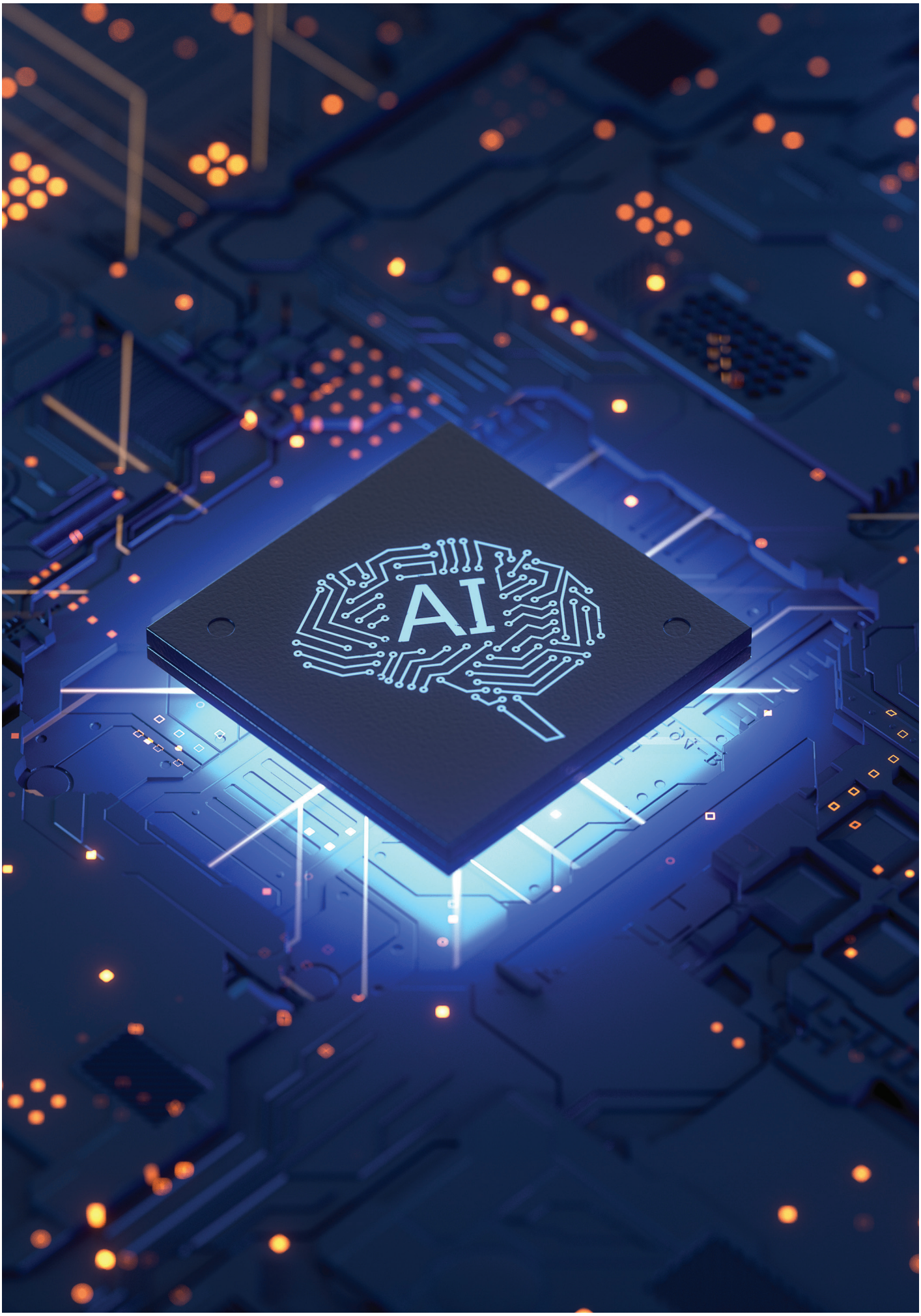
The Generative AI risk management framework rests upon a solid foundation comprising four key layers:

Business values layer: It evaluates potential risks and benefits from AI implementation, aligning projects with overarching strategic objectives, financial robustness, reputation management, and competitive edge.

Governance layer: Governance of Generative AI involves managing and overseeing its application across people, processes, and technology to ensure its responsible, secure, and ethical use. Effective governance of Generative AI requires a multidisciplinary approach involving collaboration between different teams and stakeholders. It should be an ongoing process that evolves with advancements in Generative AI technology and dynamic in nature to keep up with societal norms and regulation changes.

Five pillars: It encompass adoption, maintenance, scaling, customisation, and decommissioning, offering a comprehensive roadmap for navigating the complete Generative AI lifecycle and proactively identifying and mitigating risks at each stage.

Risk control layer: The culminating risk control layer bolsters the framework with its paramount role in ensuring that AI technologies harmonise with data security, privacy imperatives, and regulatory compliance, extending from established principles like the CIA triad to encompass the full spectrum of privacy considerations and adherence to pertinent regulations.


AI

Way forward

Some key cyber questions to help you assess your organisation's readiness for a secure, private and ethical use of Generative AI. Questions we want to leave you with:

01 What are your key business use cases to make this programme a success?

02 Do you have the right cyber investments (tools, technologies, processes, and skillset) in your strategic roadmap?

03 Is there a plan to ensure that your Generative AI tools do not threaten your organisation's end users and customers?

04 How do you ensure your Generative AI tools are not using your sensitive data for training?

05 Do you have a process in place to ensure sensitive data is not used without the right controls?

06 Do you have policies that ensure the security of your Generative AI models?

07 Do you have a Security Operations Center (SOC) to monitor threats in your Generative AI landscape?

08 How does your business ensure that only authorised users can access Generative AI tools, models, infrastructure, and data?

09 Will your Generative AI be used or built by a third party, and do you need to re-assess and re-look at your current third-party risk programmes?

10 Are you prepared to use Generative AI with your organisation's privacy and confidential controls (including consent mechanism and data sanity)?

Conclusion

Generative AI has immense possibilities regarding content, which can help reduce the effort and bring out efficiencies in the system. Generative AI has applications across the ecosystem, affecting individuals, organisations, and society alike.

While we celebrate this quantum leap in technological advancement, similar to any technology, the cybersecurity perspective that needs to be considered would enhance the scale and application of Generative AI. The risks largely lie in key areas, i.e. Generative AI models, applications, infrastructure, people, data, and the training and testing methodologies.

These risks cut across the ecosystem's foundation, bringing in human capital, technology, and industry processes. While the potential for Generative AI is undeniable, what will make it a transformational force is balancing the risk and bringing in the right controls for the global scale of adoption.

Generative AI will create immense growth opportunities in key areas, such as intelligent IT, products, and operations. The next decade will be when AI will become mainstream and further enhance human potential and growth.

Connect with us

Anthony Crasto
President, Risk Advisory
Deloitte India
acrasto@deloitte.com

Tarun Kaura
Leader – Cyber Advisory,
Risk Advisory, Deloitte India
tkaura@deloitte.com

Dr. Vikram Venkateswaran
Partner, Risk Advisory
Deloitte India
vikramv@deloitte.com

Abhijit Katkar
Partner, Risk Advisory
Deloitte India
akatkar@deloitte.com

Praveen Sasidharan
Partner, Risk Advisory
Deloitte India
psasidharan@deloitte.com

Contributors

David George
Rajat Kothari
Vivekchandran N V
Titas Nath

Deloitte.