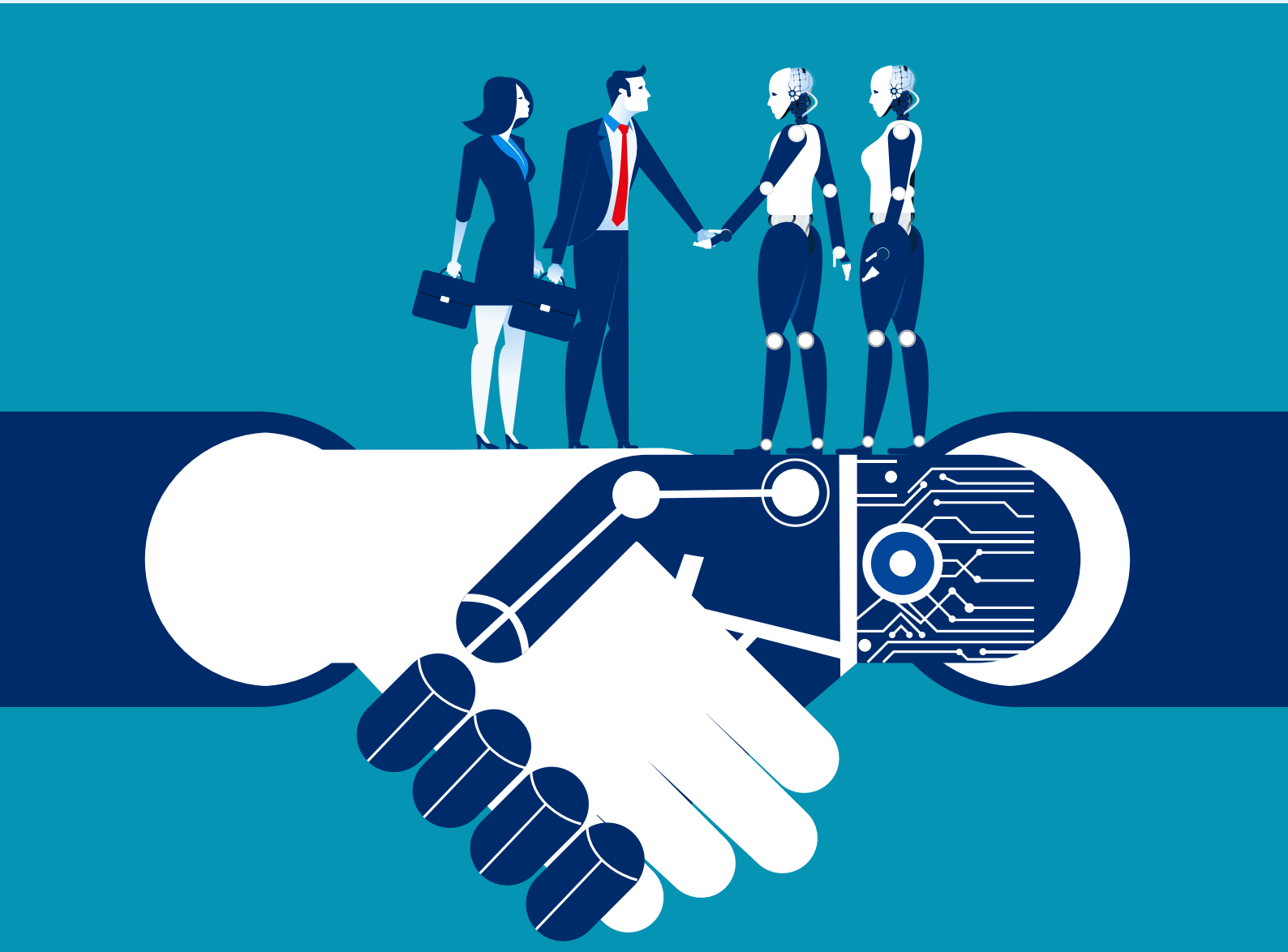
DSCI
PROMOTING DATA PROTECTION

MITIGATING SECURITY & PRIVACY RISKS

A Guide to Enterprise Use of **Generative AI**

March 2024



The report is supported by Microsoft Corporation (India) Private Limited.

CONTENTS

OBJECTIVE

The intent of this report is multi-fold. First, the report delves into the fundamentals of generative AI and how it functions. Second, it examines the current landscape of use of generative AI in an enterprise context against the privacy and security risks raised by the proliferation of generative AI adoption in the absence of established governance frameworks. Finally, it suggests some interventions in organizational policies, governance frameworks, and processes that could help mitigate the privacy and security risks identified.

The scope of this report is limited to examining the privacy and security implications of the rapid increase in generative AI adoption by enterprises. While issues pertaining to the development of **regulatory strategies, ethics in AI development and adoption,** and considerations of **user harms** are just as pertinent.

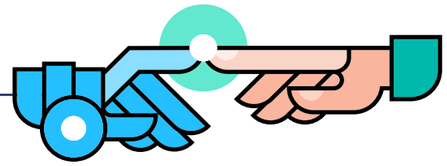
This report narrows its focus on data protection and cybersecurity considerations for enterprises and businesses as both developers and users of generative AI models.

The initial section of the document provides the requisite context to understand historical evolution of artificial intelligence models and traces the technological developments from probabilistic and discriminatory models to the emergence of models capable of generating novel outputs.

Part I focuses on cybersecurity risks and enterprise-level process and governance interventions that may help mitigate these risks. It examines the cybersecurity risks for enterprises leveraging generative AI tools and provides a classification of vulnerabilities in large language models based on their intrinsicality to the nature of the technology. This part concludes with insights for enterprises to effectively manage these vulnerabilities and aims to equip cybersecurity leaders and their teams with the relevant insights to mitigate these risks.

Part II focuses on the data protection and privacy considerations associated with the Generative AI tools and systems. This portion of the document first evaluates the manner and extent to which existing data protection norms and regulations apply to generative AI technologies. It also intends to provide an insight into the global data protection regulatory developments centering around generative AI. This section ultimately identifies the specific data governance and protection challenges that emerge from the increasing adoption of generative AI models in enterprises and provides some strategic directional guidance for businesses to mitigate these risks.

BACKGROUND



In recent years, the field of Artificial Intelligence (AI) has undergone a remarkable transformation with the emergence of Generative AI. Unlike traditional AI models that are deterministic and follow strict programmed rules, generative models leverage the inherent patterns in data to replicate and adapt them creatively. This branch of AI represents a significant paradigm shift, moving away from passive consumption towards active creation, allowing machines to generate original content.

Generative AI's versatility extends to creating coherent text, composing music, designing graphics, simulating human speech, crafting 3D objects, and formulating scientific hypotheses. Its applications span diverse sectors, offering new avenues for business operations. Yet, this innovation also amplifies existing cybersecurity risks by lowering technical barriers, potentially empowering less-skilled individuals to experiment with sophisticated cyber-attack techniques. Alongside its promise, generative AI presents a spectrum of risks that demand careful consideration.

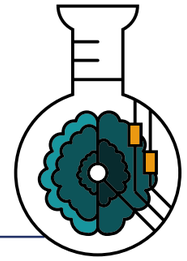
Generative AI's capabilities can scale up reconnaissance to an unimaginable level and scope. The planetary-scale collection and processing of data can provide rapid, multidimensional, and multi-planar insights useful for launching future campaigns. The attack surface will grow exponentially due to Large Language Models (LLMs). It also raises some data protection related concerns and risks. The need for extensive datasets, often containing personal information, heightens the risk of data exposure and unauthorized access. The complex value chain of stakeholders and the diversity of use cases where such models can be utilized, also make it difficult to define concrete measures to adhere to data protection principles. At the output generation layer, biased training data may perpetuate discrimination against individuals and the threat of hallucinations in a model can lead to factually inaccurate inferences.

These cybersecurity and privacy implications underscore the importance of thoughtful regulation and proactive measures in leveraging generative AI responsibly.

Generative AI refers to AI techniques that learn a representation of artifacts from data, and use it to generate brand-new, unique artifacts that resemble but don't repeat the original data.[1]

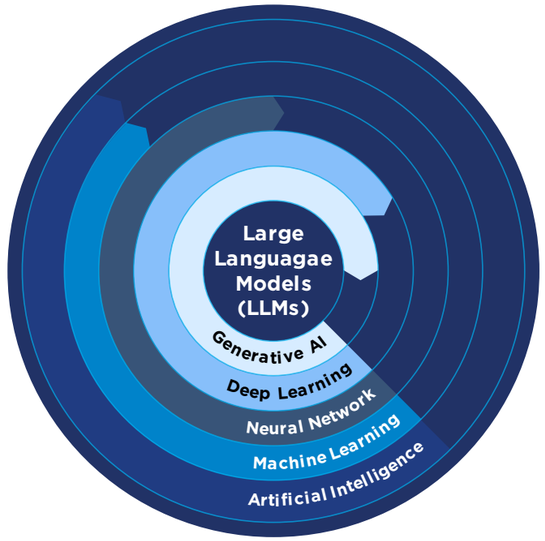
HISTORICAL EVOLUTION:

From Probabilistic to Generative Models



AI encompasses the capability of computers to mimic human behaviors. Within AI, Machine Learning (ML) employs mathematical methods to enhance computer performance through data-driven learning. Deep Learning (DL) utilizes layered neural networks for computation, aiming to enable computers to learn and make intelligent decisions autonomously. This advancement in AI has revolutionized the ability of machines to perform complex tasks previously reserved for humans.

AI's journey traces back to machine learning attempts in the 1950s and 1960s, with Alan Turing's milestone research, 'Computing machinery and intelligence; which focused on the question, "Can machines think?" [2]



Limited resources slowed progress until the 1990s and 2000s, when advanced hardware emerged. Generative AI surfaced with neural networks, emulating the human brain's interconnected 'neurons,' recognizing patterns without explicit programming. Ian Goodfellow's Generative Adversarial Network (GAN) in 2014 sparked the field's growth, joined by models like Variational Autoencoders (VAEs) and Recurrent Neural Networks (RNNs), showcasing their content generation abilities. From basic language models to sophisticated content generators, generative AI evolved into an innovation powerhouse. Gartner predicts over 100 million people will be engaging robo-colleagues by 2026, thus transforming enterprise work.[3]

The roadmap below examines the development of AI over the last few decades and delves into the advancements which culminated into the age of generative AI that we are witnessing presently.[4]

Beginnings & Milestones

1960

Eliza showcasing early AI prowess in natural language processing by simulating conversations akin to a psychotherapist, showcasing substantial progress in interaction and comprehension.

1980

Procedural Content Generation in Rogue, by Michael Toy and Glenn Wichman, revolutionized gaming with procedural content generation, creating dynamic game levels on the fly. This pioneering technique set the stage for AI's role in generating creative content in the future.

The Rise of Neural Networks

1986

Michael Irwin Jordan's advancements in Recurrent Neural Networks (RNNs) marked a pivotal moment in sequence modelling, establishing the foundation for contemporary neural network structures.

2000

Researchers from the University of Montreal introduced a Neural Probabilistic Language Model, employing feed-forward neural networks to model language, thereby expanding avenues for language comprehension and generation.

Decade of Transformation: 2010-2020

2011

Siri, Apple's voice-activated personal assistant, showcased the capabilities of Generative AI by generating responses and executing tasks through voice commands.

2012

Fei-Fei Li's ImageNet database and Alex Krizhevsky's AlexNet architecture were pivotal in advancing visual object recognition and image generation, propelling significant strides in this field.

2013

Google's Tomas Mikolov introduced Word2Vec, enabling automated detection of semantic connections among words, greatly improving language comprehension and generation capabilities.

2014

The advent of diffusion models involves integrating noise into training data and subsequently reversing the process to reconstruct and restore the original data.

2015

Ian Goodfellow's Generative Adversarial Networks (GANs) and Diederik Kingma's Variational Autoencoders (VAEs) pioneered pathways for creating lifelike images, videos, and text, pushing the boundaries of generative capabilities.

2017

The proposal of the transformer architecture revolutionized deep learning.

2018

OpenAI introduced the groundbreaking Generative Pre-trained Transformers (GPT), a significant leap in large language models.

Advancements in Generative AI Today

2021

The launch of DALL-E, an AI platform designed for crafting and modifying distinctive artworks and photorealistic images, took place.

2022

Novel AI Image Generation Tools Unveiled - Stable Diffusion, an open-source tool, and proprietary Midjourney AI were introduced. Stable Diffusion, enabling automatic image content creation from text, was released by researchers.

2023

In March 2023, the launch of GPT-4 marked a milestone, boasting the ability to produce lengthy texts spanning up to 25,000 words. The technology landscape witnessed significant shifts as Microsoft integrated ChatGPT into its Bing search engine.

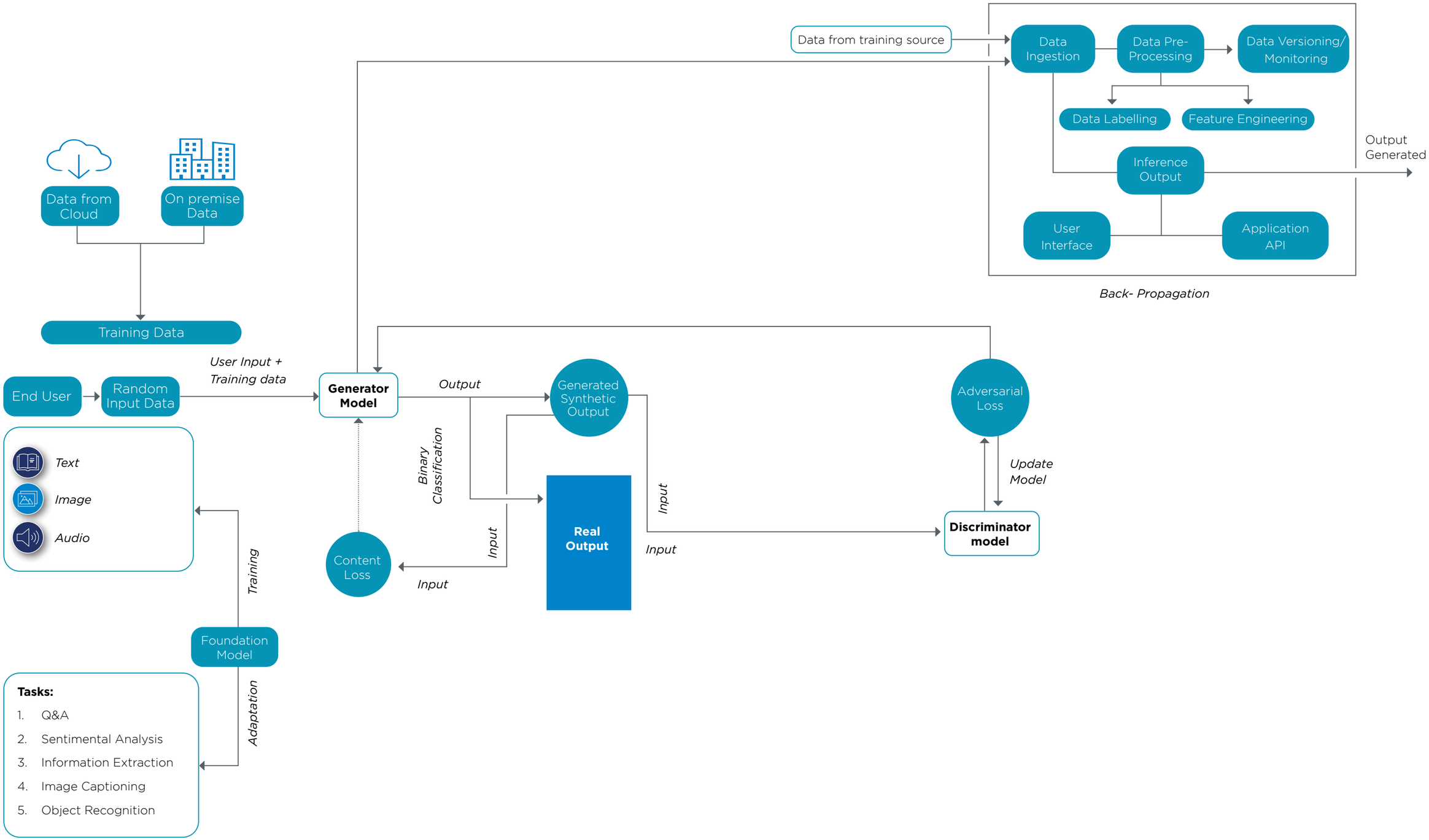
2025

By 2025, the horizon will be filled with increasingly sophisticated AI-driven tools tailored for marketers, facilitating the swift recognition of consumer behavior trends and patterns.

2030

AI has already made impactful inroads into various sectors such as healthcare, finance, and transportation. Looking ahead to 2030, its presence will be even more pervasive, seamlessly weaving into our daily lives through smart homes, self-driving vehicles, and intelligent personal assistants.

The Architecture of Generative AI Technology



The user interacts with the model by providing prompts, and the model leverages data from both cloud and on-premises sources. This data can be diverse, including text, images, speech, structured data, or signals, and is used for various tasks like question answering, language generation, sentiment analysis, and more.

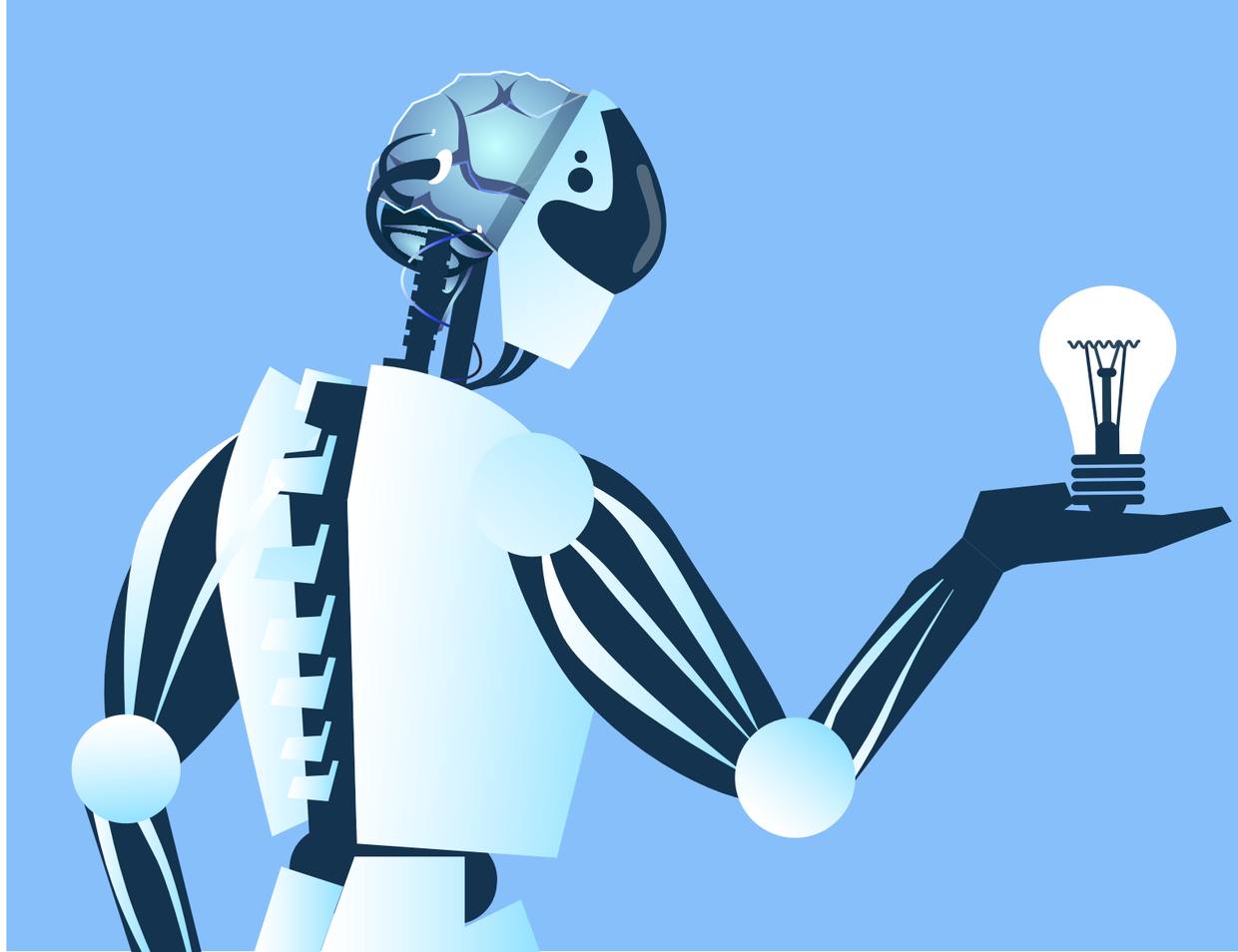
The model comprises of two neural networks, **Generator and Discriminator,** engaged in an adversarial setup. This setup forms a zero-sum game, where the success of one network implies a loss for the other. The generator takes combined data from user prompts and training sources, creating synthetic output. This output is then fed into the discriminator, which acts as a binary classifier, distinguishing between real and synthetic sample The network continually strives to improve; the generator aims to create convincing fakes, while the discriminator sharpens its ability to differentiate real from fake.

Implemented typically as Convolutional Neural Networks (CNNs), both the generator and discriminator often focus on image-related tasks. The goal is for the generator to craft samples that are indistinguishable from real ones, challenging not only the discriminator but also human observers. This competitive cycle updates each network, aiming for continuous performance enhancement.

Within the generator model, the training data undergoes processes like data ingestion, pre-processing, feature engineering, and monitoring. The inference output from this process is then fed into the discriminator model, where outputs undergo methods such as model serving, monitoring, management, and evaluation through repositories. The overall goal is to enhance the performance of the GenAI model in generating realistic outputs and effectively discerning between real and generated data.

Generative AI Workflow

PART 1 STRENGTHENING CYBERSECURITY FOR GENERATIVE AI



Exploring Cybersecurity Risks*

By 2026, over 80% of enterprises are projected to embrace GenAI, compared to a mere 5% in 2023

The rapid adoption of LLMs and Generative AI (GenAI) is transforming businesses, promising significant advancements. However, unmanaged adoption of these technologies can pose serious cybersecurity risks. As their capabilities in cybersecurity evolve, a two-fold approach is crucial; first, understanding the inherent risks and second, developing robust strategies to mitigate them.

By 2026, over 80% of enterprises are projected to embrace GenAI, compared to a mere 5% in 2023[5]. This rapid growth highlights the growing importance of Gen AI, but also underscores the need to address potential misuse. Malicious actors could weaponize LLMs for crafting sophisticated cyber-attacks, spreading misinformation, or even gaining unauthorized access to sensitive data.

To navigate this complex landscape, we must delve deeper into the vulnerabilities associated with LLMs. These range from hardware and software vulnerabilities to user-level threats like misinformation and fraud. Recognizing these risks and proactively implementing strong cybersecurity measures is essential.

Hardware-Level Threats

Hardware attacks typically require physical access to devices. However, since LLMs cannot directly access physical devices, they can only interact with information associated with the hardware. Nevertheless, LLMs can inadvertently enable side-channel attacks, which involve analyzing unintentional information leaks from physical systems to infer secret information, such as cryptographic keys.

Operating System-Level Vulnerabilities

LLMs operate at a high level of abstraction and primarily handle text-based input and output, lacking the necessary low-level system access for executing OS-level attacks. Nonetheless, they can be leveraged to analyze information obtained from operating systems, potentially aiding in the execution of such attacks.

Software-Level Exploits

LLMs have been employed in software attacks, for instance, creating malware. Malicious developers can utilize LLMs like ChatGPT to distribute undetected malicious software or create various types of malwares, including ransomware, worms, keyloggers, and fileless malware.

Network-Level Risks

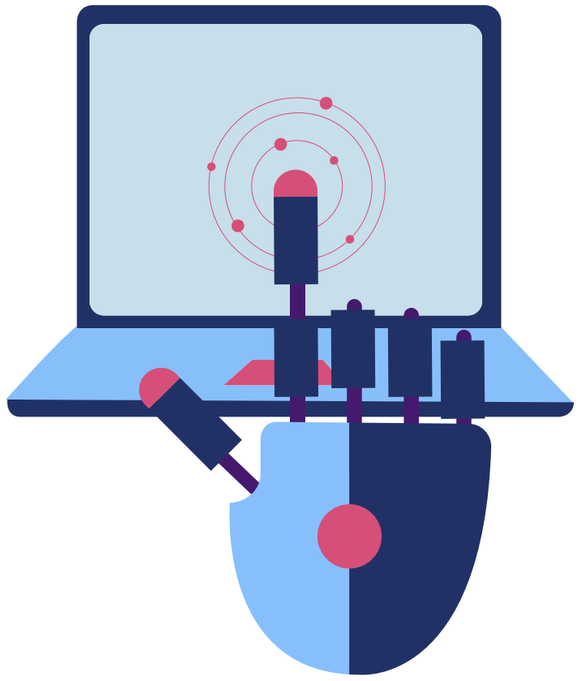
LLMs can be utilized to initiate network attacks, such as phishing. Modifying inputs to LLMs like ChatGPT can influence the content of generated emails, making them more convincing.

User-Level Threats

LLMs can generate highly convincing yet deceptive content, posing risks in various user interactions:

- **Misinformation:** Synthetic content generated by LLMs raises concerns about the integrity of online information. For example, deepfakes are fabricated videos and audios that manipulate public opinion, along with fake news and social media bots.
- **Social Engineering:** Well-trained LLMs can infer personal attributes from text and extract sensitive information from seemingly innocuous queries. For example, phishing emails, impersonation (LLMs mimic writing styles to impersonate real people online for malicious purposes), and spam and comment flooding (overwhelming platforms with irrelevant content drowns out legitimate communication).
- **Fraud:** Tools like FraudGPT and WormGPT operate similarly to ChatGPT but lack safety controls and are sold on the dark web. They enable cybercriminals to create fraudulent emails, plan attacks, and execute Business Email Compromise (BEC) attacks.

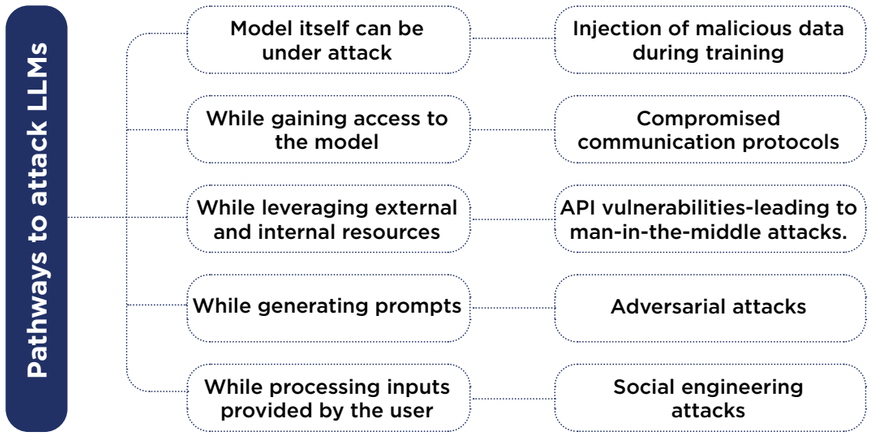
The capabilities of LLMs to produce realistic text and mimic human behavior pose challenges to traditional defense mechanisms, such as CAPTCHA challenges, and increase the risk of fingerprinting attacks. These cybersecurity risks highlight the importance of implementing robust measures to mitigate potential threats associated with generative AI technologies.



Generative AI spans across diverse content generation technologies like images, videos, music, and text. However, LLMs concentrate solely on text generation and comprehension, thus serving as a subset of generative AI specialized in language tasks. Common cybersecurity risks observed in LLMs are also inherent in generative AI. Both share analogous traits and vulnerabilities, exposing them to comparable cybersecurity risks.

Understanding Vulnerabilities and Threats

Now that we've explored the broader cybersecurity risks associated with Generative AI, let's shift our focus to understand the vulnerabilities and threats specific to LLMs. These vulnerabilities pertain to weaknesses inherent to LLMs, such as susceptibility to adversarial attacks or unintended biases in generated content. It's important to distinguish these model-specific concerns from the broader spectrum of cybersecurity risks arising from the deployment and utilization of various generative AI technologies.



AI Inherent Vulnerabilities and Threats

It refers to vulnerabilities and threats inherent to LLMs. For instance, attackers might manipulate input data to produce incorrect or undesired outputs from the LLM.

- **Adversarial Attacks:** These involve intentional manipulation or deception of machine learning models, exploiting vulnerabilities in the model's behavior for malicious purposes.
- **Data Poisoning:** Attackers influence the training process by inserting malicious data into the training dataset, compromising pre-trained models through methods such as injecting poisoned content.
- **Backdoor Attacks:** Malicious manipulation of training data and model processing creates vulnerabilities where hidden backdoors can be embedded, altering specific behaviors or responses when triggered.[i]

[i] *Both backdoor attacks and data poisoning attacks involve the manipulation of machine learning models, which can encompass the manipulation of inputs. However, the crucial difference lies in the fact that backdoor attacks concentrate on embedding concealed triggers within the model to influence particular behaviors or responses upon trigger activation.*

- **Inference Attacks:** Adversaries attempt to extract sensitive information about a machine learning model or its training data by making specific queries or observations, exploiting unintended information leakage from responses.[ii]
- **Inference Attacks:** Adversaries attempt to extract sensitive information about a machine learning model or its training data by making specific queries or observations, exploiting unintended information leakage from the responses.
- **Extraction Attacks:** Adversaries extract sensitive information or insights directly from machine learning models or their associated data, aiming to acquire specific resources or confidential information.[iii]
- **Bias and Unfairness Exploitation:** This concerns prejudiced outcomes or discriminatory behaviors exhibited by LLMs, which have raised ethical and societal concerns.
- **Instruction Tuning:** LLMs are fine-tuned for specific tasks by providing explicit instructions or examples, which can be exploited to reveal vulnerabilities or limitations.
- **Jailbreaking:** Bypassing security features to enable responses to restricted or unsafe queries, unlocking capabilities typically limited by safety protocols.
- **Prompt Injection:** Manipulating LLM's behavior to elicit unexpected and potentially harmful responses by crafting input prompts to bypass safeguards.

Non-AI Inherent Vulnerabilities and Threats

It encompasses external threats and new vulnerabilities not typically observed or investigated in traditional AI models.

- **Remote Code Execution (RCE):** RCE attacks, though not directly targeting LLMs, pose a significant threat. Exploiting vulnerabilities in web services or platforms where LLMs are integrated, attackers could execute arbitrary code remotely, compromising LLM environments.
- **Side Channel Vulnerabilities:** While LLMs don't typically leak information through traditional side channels, they are vulnerable to certain side-channel attacks in practical scenarios. Privacy side-channel attacks, for instance, exploit system-level components to extract private information at a higher rate.
- **Insecure Plugins:** Third-party plugins, while enhancing LLM functionality, introduce security risks. Concerns include the potential for stealing sensitive data or executing malicious code. OAuth usage in plugins further amplifies these vulnerabilities.

[ii] *In Attribute Inference Attacks, the attacker seeks to infer sensitive or personal information about individuals or entities by scrutinizing the behavior or responses of machine learning models. Membership inference Attacks aim to ascertain whether a data record was included in a model's training dataset, utilizing either white-box or black-box access to the model and the specific data record.*

[iii] *Extraction attacks strive to directly obtain specific resources like model gradients, training data, or confidential information. In contrast, inference attacks aim to glean knowledge or insights about the characteristics of the model or data, typically by observing the responses or behavior of the model. This encompasses various techniques such as model theft attacks, gradient leakage, and extraction of training data.*

Navigating Unmanaged Adoption: Generative AI in Enterprise Context

Nearly half (48%) of the organizations do not have specific guidelines and/or policies put into effect yet for responsible AI.[6]

Generative AI serves as a catalyst for digital transformation, empowering businesses to make data-driven decisions, personalize customer experiences, and streamline operations. However, alongside harnessing the benefits of generative AI, understanding the cybersecurity landscape is essential to ensure the resilience and security of digital endeavors.

Let's explore the taxonomy of LLMs within the enterprise context.

- **Third-party LLMs:** These are services provided by external entities for end-user consumption. Examples include platforms like OpenAI, where businesses can leverage pre-trained models to enhance their operations.
- **Consumer LLMs:** Organizations that develop generative AI as a service fall into this category. They build and offer generative AI capabilities to consumers or other businesses as part of their products or services.
- **Employee LLMs:** These are deployed internally within an organization and are tailored to specific departments or functions. Employee LLMs utilize organization-specific data to facilitate seamless intra-organizational communication and workflow optimization.

Building Cyber-Resilient Enterprises: Key Security Considerations for Generative AI Adoption

Enterprises must prioritize the adoption of a proactive and comprehensive approach to cybersecurity to safeguard their digital assets, protect sensitive information, and maintain the trust of their stakeholders.

Security governance has come to the point that cyber issues must be managed in real time, while also guaranteeing productivity and efficiency and assuring compliance with stringent regulations.

- **Robust Governance Frameworks:** Implement clear and comprehensive governance frameworks that define acceptable use policies for generative AI models across the organization. This framework should address:
  - **Authorized Users:** Clearly define which roles and departments are permitted to use generative models, ensuring responsible access and usage.
  - **Automation Scope:** Specify the specific processes and tasks where generative AI can be used for automation or enhancement, ensuring alignment with organizational goals and risk tolerance.
  - **Data Access Controls:** Establish clear guidelines on which internal applications and data sets are accessible to these models, and how they can be used. This minimizes the risk of unauthorized access or misuse.
  - **Clear Disclosures:** Establish a policy requiring employees to disclose when internal or external work products are created in whole or part by generative AI tools. This transparency helps build trust, identify potential risks, and ensure compliance.
- **Zero-Trust Platforms:** Implement zero-trust platforms with anomaly detection to proactively identify and mitigate threats, lowering the risk of breaches.
- **Prompt Scrutiny:** Carefully review prompts used in generative AI platforms to prevent accidental disclosure of intellectual property or sensitive information.
- **Enhanced Controls:** Deploy encryption and access controls to secure data, preventing unauthorized access and breaches. Regular audits and risk assessments ensure secure endpoints and compliance with security protocols.
- **Employee Training:** Provide comprehensive training on responsible AI use, empowering employees to understand and mitigate data privacy and security risks. Encourage critical evaluation of outputs and adherence to best practices.
- **Regulatory Compliance:** Stay updated on data privacy regulations like DPDPA, GDPR, CPRA, and industry-specific requirements.
- **Cybersecurity Investments:**
  - **Robust Tools:** Invest in robust cybersecurity solutions to protect generative AI models and technologies from cyber threats. Configure network security tools effectively, considering generative AI models as part of the attack surface for data protection.
  - **Privacy-Enhancing Technologies (PETs):** Explore traditional PETs like zero-knowledge proofs, differential privacy, and federated learning to address privacy challenges posed by LLMs. Consider innovative PETs techniques as well to tackle these concerns effectively.
- **Prefer First-Party Data; Otherwise, Responsibly Source Third-Party Data:** Knowing the origin of your input data is crucial for using generative models in business. Prioritize using your organization's data whenever possible. Ensure proper authorization if you need

to use third-party data. Investigate how suppliers source their data and avoid those lacking transparency to mitigate legal risks.

Technical Controls

As discussed earlier, the widespread adoption of LLMs and generative AI presents exciting possibilities but also raises critical security concerns. Unlike traditional software, these evolving models are susceptible to manipulation and misuse.

This section further explores crucial technical controls that organizations can employ to safeguard generative AI, thus promoting trust and upholding responsible utilization of these potent technologies.

- **ISO 27032:**

The principles outlined in ISO 27032 can be effectively applied to LLM security through the following practices:

i **Risk Assessment:** Conduct a thorough evaluation to identify potential vulnerabilities and threats specific to generative AI models, data, and workflows. Utilize the guidance to develop a detailed risk management plan that outlines strategies to mitigate these identified risks.

ii **Security Policies:** Develop clear and concise security policies governing the use of LLMs and generative AI, ensuring they align with the organization's overall risk management strategy. The framework can guide the development of these policies, ensuring they address the unique security considerations of LLMs.

iii **Effective Incident Management:** Establish a well-defined incident management plan to effectively respond to security incidents involving LLMs. It provides guidance on developing and implementing such a plan, minimizing the impact of potential attacks and enabling a swift recovery process.

- **MITRE ATT & CK and ATLAS:**

It offers valuable resources for understanding and mitigating threats specific to machine learning systems. By mapping the LLM's security strategy to these frameworks, organizations can identify areas where existing security processes, such as API security standards, adequately address potential vulnerabilities. This mapping exercise also highlights gaps in security controls, allowing organizations to prioritize and implement necessary safeguards.[7]

- **THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)**

The National Institute of Standards and Technology (NIST) has developed a classification and vocabulary for adversarial machine learning threats and countermeasures.[8] This resource is intended for those involved in various aspects of AI, including design, development, deployment, evaluation, and governance. The classification system is structured across five key aspects: the type of AI system (Predictive or Generative), the phase of the machine learning lifecycle where attacks take place, the objectives of the attacker, the capabilities of the attacker, and the attacker's knowledge of the learning process.

Lastly, national and international standard-setting organizations should prioritize the development of robust evaluation criteria for LLMs, with a focus on ensuring the diversity of corpora used in training, transparency in model operations, accuracy of outputs, and the implementation of rigorous data security measures.

Cybersecurity Leaders Toolkit

The emergence of generative AI demands a swift and decisive response from cybersecurity leaders. It's no longer a distant threat, but a pressing concern requiring immediate attention. Here's a proactive approach to mitigate risks.

- **Assess AI Exposure:**

Bring together cybersecurity, technology, data, and operations leaders for discussions at the board level regarding emerging risks associated with generative AI. Address potential vulnerabilities that could lead to unauthorized access or exposure of sensitive data due to adversarial AI techniques.

- **Secure the AI Pipeline:**

Prioritize the security and encryption of data used in training and fine-tuning AI models. Continuous monitoring for vulnerabilities, malware, and data corruption throughout the model development process and post-deployment to detect and mitigate AI-specific attacks.

- **Invest in Specialized Defenses:**

While existing security measures can extend to protect AI infrastructure and data, combating adversarial attacks on AI models requires innovative defense strategies.

Assess the necessity of AI adoption and its suitability for addressing specific problems.

Checklist for AI Adoption

- Evaluate the impact on privacy and confidentiality and ensure compliance with governance and contingency measures.
- Determine the readiness of the organization for AI implementation, including in-house expertise and data quality.
- Establish criteria for vetting AI providers and integrating AI solutions into existing enterprise workflows.

Checklist for Generative AI Solution Providers

Could the solution provider furnish detailed insights into their data management protocols, specifically regarding how they handle data access and transmission beyond our organizational boundaries? It's crucial to understand how they safeguard organization's data, especially when it's accessed or transferred beyond our control.

What security measures and performance metrics does the solution provider adhere to? Are there any peer-reviewed assessments available that attest to the effectiveness of their security protocols? Furthermore, how seamlessly can their solutions integrate with third-party security tools? This information is vital for ensuring the compatibility and effectiveness of their security measures within the organization's existing infrastructure.

Equipping Teams to Manage Generative AI: A Proactive Approach

- **Mapping the AI Landscape:**
  - i **Unveiling Usage:** Utilize audits, surveys, and endpoint monitoring tools to uncover who's using AI and for what purposes.
  - ii **Demystifying the Need:** Understand the driving force behind the demand for AI tools and their true value to your organization.
- **Weighing Risks and Rewards:**
  - i **Business Impact Assessment:** Analyze each AI use case, meticulously weighing its benefits against potential security implications and privacy concerns.
  - ii **Risk-Benefit Equation:** Strike a balance between the advantages and potential risks, adjusting data access permissions as needed.
- **Building a Governance Framework:**
  - i **Policy Compliance:** Ensure that AI practices strictly adhere to company policies and establish risk tolerance levels.
  - ii **Controlled Experimentation:** Create sandboxes for testing AI technologies and mitigating potential risks before real-time deployment.
  - iii **Supervised Exploration:** Encourage supervised exploration of new AI use cases with controlled testing and rollout strategies.
- **Keeping a Watchful Eye:** Maintain vigilant monitoring of AI outputs, especially during the initial deployment phase.
- **Collaborative Data Classification:** Partner with CISOs, tech teams, and risk management experts to effectively classify data. Identify and isolate highly sensitive data, restricting access for AI tools.
- **Data Integrity First:** Implement robust data classification practices to enhance security and safeguard data integrity.
- **Validating the Code:**
  - i **Security Scanners:** Rigorously scan AI-generated coding outputs for any potential security vulnerabilities.
  - ii **Validation Processes**: Implement robust validation processes to effectively address and mitigate potential security threats.

Considerations for Developers: Security Guardrails

Data Preparation

Corpora Cleaning: LLMs are molded by their training corpora, which dictate their behavior, concepts, and data distributions. Hence, the quality of training corpora profoundly impacts the safety of LLMs. However, raw corpora sourced from the web often contain issues of fairness, privacy, and credibility, necessitating careful curation. The figure below illustrates the data preprocessing steps in generative AI, each serving a specific purpose such as ensuring accurate language comprehension, filtering harmful content, addressing biases, safeguarding user privacy, and optimizing dataset efficiency for training.

Research Priorities: LLM developers should prioritize research in areas such as building easily maintainable models and evaluating model fitness for specific tasks. Robust training approaches, including adversarial training and robust finetuning, bolster model resilience against attacks and enhance defense mechanisms.

Inference Phase

Instruction Processing (Pre-Processing): Pre-processing involves sanitizing user instructions to eliminate potentially malicious content or contexts, minimizing the risk of encountering adversarial inputs.

Malicious Detection (In-Processing): Detecting anomalous patterns or indications of malicious intent within LLM computational processes enhances sensitivity and specificity in identifying harmful inputs.

Generation Processing (Post-Processing)*:* Evaluating the properties of generated outputs, including their potential for harm, allows for adjustments to mitigate identified risks before presenting responses to users, ensuring safety and appropriateness.

Fine-Tuning

Fine-tuning LLMs for specific tasks relies on developer judgment, potentially introducing biases and inefficiencies. In addition, democratizing fine-tuning poses risks of malicious input.

Robust documentation of the fine-tuning process, along with technical safeguards like data augmentation and transfer learning, can help mitigate these concerns and promote transparency and accountability.

Lack of Standardized Evaluation

There's no universal standard for assessing LLM outputs, making it difficult for third parties to evaluate performance objectively. Developers often conduct tests, but inconsistent disclosure of methodologies and results hinder transparency.

Establishing a standardized evaluation framework is crucial to promote transparency by:

- i Disclosing evaluation methods and results.
- ii Encouraging independent evaluations.
- iii Fostering collaboration within the AI community to share best practices.

Security Through Human Oversight

Integrating Human-in-the-Loop (HITL) review strengthens security by leveraging human expertise. Humans can:

i Detect and rectify errors.

ii Alleviate biases.

iii Moderate unsuitable content.

iv Ensure legal compliance.

v Manage contextual complexities.

vi Respond promptly to evolving situations.

Responsible Practices of LLM and App Developers

Design for Specific Purposes: LLMs should be designed and evaluated for specific purposes rather than assumed to be universally applicable. Collaboration between app and LLM developers ensures responsible development and deployment.

Transparency and Limitations: LLM developers must be transparent about technology limitations, especially when discussing with app developers. App developers should avoid using LLMs for unsuitable tasks. They need to understand that LLMs may not accurately represent up-to-date language patterns or understand beyond language modeling.

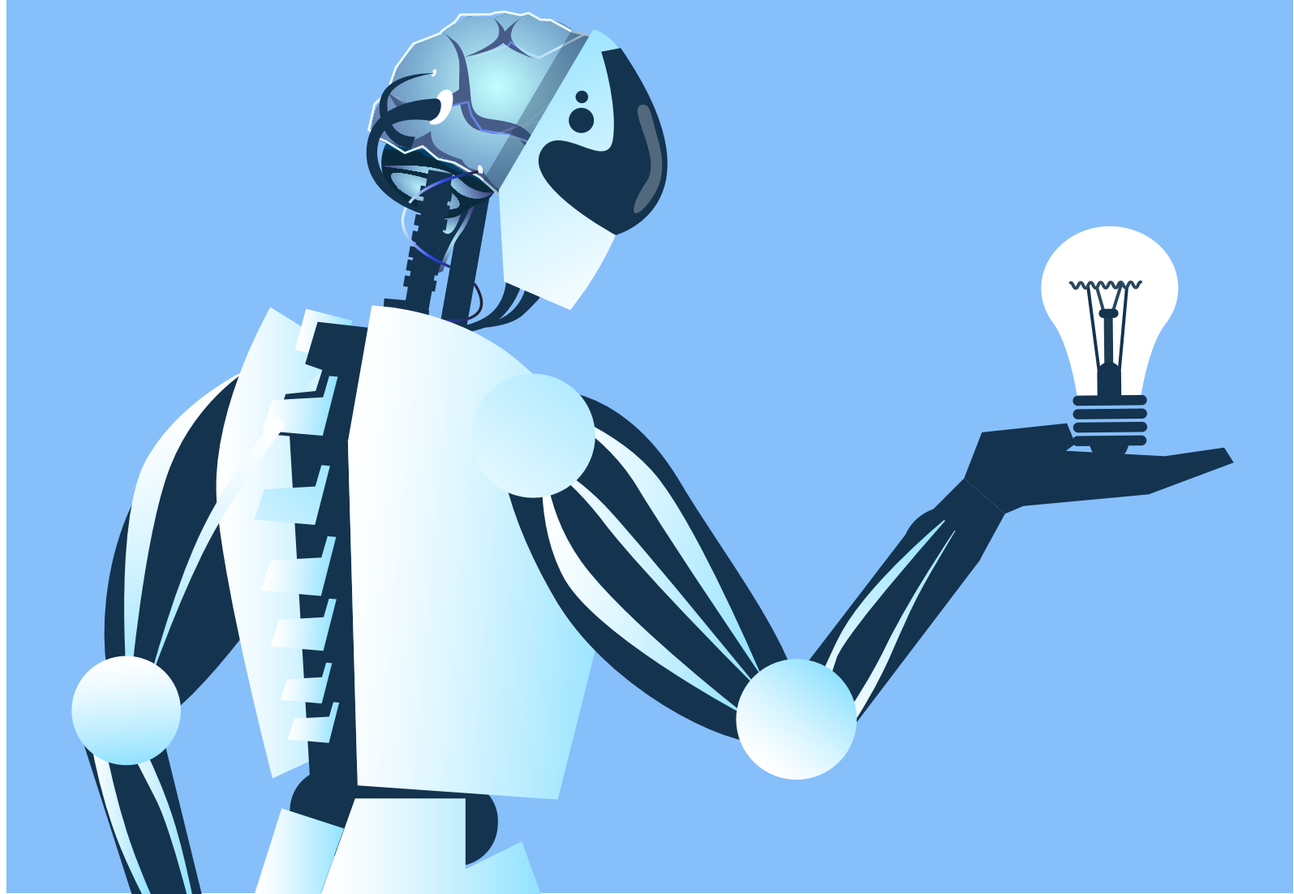
LLMs are trained on massive datasets like Common Crawl, which represent only a portion of websites and can contain inherent biases. This poses a risk of biased or harmful outputs without additional, carefully curated training.

Tokenization, the process of breaking down text into units for analysis, can also influence LLM performance in unforeseen ways due to the chosen algorithm. Different approaches, like word-level or component-level tokenization, can have varying impacts.

Ensuring the security of AI models is crucial in today's rapidly evolving AI development landscape. By incorporating practices such as thorough threat modeling, secure coding, robust DevSecOps, data encryption, privacy engineering, secure computing, regular updates, strong authentication, adversarial training, and transparent model interpretation, developers can instill reliability and trustworthiness in their AI models.

As AI technology advances, a proactive stance on security becomes essential to tackle emerging threats and foster a safer and more secure AI-driven environment.

PART II DATA PROTECTION STRATEGIES FOR GENERATIVE AI

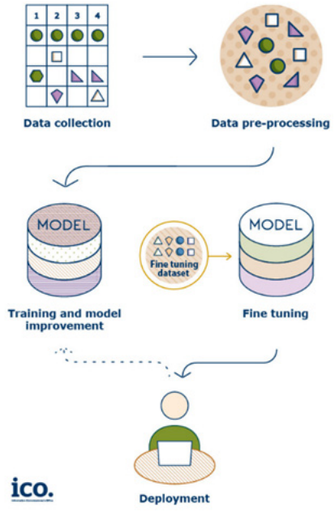


Applicability of Data Protection Regulations and Principles to Generative AI

Generative AI models are trained on massive datasets which enables these AI models to undertake highly powerful processing activities.

Use of Personal Data in Training and Development

Generative AI models are trained on massive datasets which enables these AI models to undertake highly powerful processing activities. The data pre-processing phase for developing a generative AI model may be understood to include three key elements; data collection and aggregation, data cleansing, and data scaling.[9] It is during the data collection and aggregation stage that personal data may be incorporated into the learning process.



U.K. ICO's infographic demonstrating the model development lifecycle for generative AI and the role of data at each step.[10]

For instance, the GPT-3.5 language processing model was reportedly trained on text databases including over 570 GB of data obtained from different types of written content on the internet.[11] The GPT-4 model was also trained on extensive volumes of data using both publicly available data (such as data available on the internet) and data licensed from third-party providers.[12] Technical documentation of the GPT-4 model indicates that while attempts have been made to fine-tune models to reject requests requiring access to personal data and to remove personal data from training datasets, yet the model is susceptible to attempts for identifying individuals when its capabilities are augmented with external data points.[13]

Publicly available datasets such as Common Crawl[14], or Large-Scale Artificial Intelligence Open Network[15] are also often used as part of the troves of training data required in the development phase of generative AI models. Web crawling and web scraping techniques may be leveraged to access and source data for AI training.[16]

The legality of web scraping, and the use of publicly available personal data for training and development of AI models is a question which has not been answered determinatively. Across jurisdictions, in the past few years, consultations and investigations have been initiated to determine whether web scraping would qualify as lawful processing of personal data under the applicable data protection laws and regulations.

For instance, in the European Union, United Kingdom, and Canada national data protection authorities have initiated investigations and consultations on the privacy implications of generative AI models, specifically on publicly available and consumer-facing LLMs. A brief description and timeline of these regulatory initiatives is given below.

Garante, Italy (March, 2023)

In the absence of detailed information being provided to data subjects about personal data collected from them, and a lack of clarity surrounding the legal basis on which the data was being processed to train algorithms, ChatGPT was stopped from operating in Italy temporarily.

Upon some measures being undertaken, access to the service was restored for Italian users in April 2023.

'Artificial Intelligence: stop to ChatGPT'; Garante Per La Protezione Dei Dati Personali, 'Italy restores ChatGPT after OpenAI responds to regulator'; REUTERS,

CNIL, France (April, 2023)

Upon receiving over five complaints, the French supervisory authority for data protection, CNIL, opened a formal investigation into ChatGPT.

The basis of the complaints seemed to range from lack of transparency and fairness in data processing to the inability of the data subjects to exercise their right to access their personal data.

'Governments vs. ChatGPT: Investigations around the world', DIPLO, 'Artificial Intelligence: The action plan of the CNIL', CNIL.

DPC, Ireland (July, 2023)

In July 2023, Google's Bard was made available to users in the EU after initial delay because of the Irish Data Protection Commissioner's concerns around the lack of transparency and the absence of appropriate data protection impact assessment documentation at the time.

While there is no publicly available information about a formal investigation, the European Data Protection Board's (EDPB) taskforce will continue to examine the compliance of AI chatbots with data protection regulations.

'Google delays EU launch of its AI chatbot after privacy regulator raises concerns', TECH CRUNCH, 'Google's Bard and other AI chatbots remain under privacy watch in the EU', TECH CRUNCH

ICO, U.K. (December, 2023)

In the U.K., the ICO has initiated a consultation to determine whether there is a lawful basis to use personal data scraped from the web to train generative AI models.

The U.K. ICO has also previously issued a preliminary enforcement notice to a popular social media platform for privacy risks associated with its generative AI-powered chatbot.

'Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models', ICO, 'UK Information Commissioner issues preliminary enforcement notice against Snap', ICO

OPC, Canada (December, 2023)

The Office of the Privacy Commissioner in Canada opened an investigation into ChatGPT, jointly with the country's provincial privacy authorities in May 2023.

More recently, in December 2023, the Privacy Commissioner published some guiding principles which encourage organizations to exercise caution before scraping publicly accessible personal information.

'OPC to investigate ChatGPT jointly with provincial privacy authorities', Office of the Privacy Commissioner of Canada, 'Principles for responsible, trustworthy and privacy-protective generative AI technologies', Office of the Privacy Commissioner of Canada

A timeline and brief description of data protection related regulatory initiatives in the context of generative AI services

In a joint statement, twelve members of the Global Privacy Assembly's International Enforcement Cooperation Working Group, also published their opinion on data scraping and its impact on protection of privacy rights of individuals.[17]

The independent investigations undertaken by data protection authorities and regulators from across the globe, and the joint statement mentioned above, indicate that use of personal data in the training of generative AI models, especially where such data is collected through web scraping, can have several consequences for an individual. First, in most jurisdictions publicly available or accessible personal data is still subject to data protection laws and regulations which indicates that developers must have a lawful basis to process this data, create modalities for data subjects to exercise their rights, etc.

Second, data scraping at scale could by itself constitute or increase the risk of occurrence of a data breach. Finally, the status of investigations into the use of personal data for generative AI training reveals that there is a lack of complete clarity and transparency about how personal data is used in these processes.

There are also some risks which are identifiable for an enterprise user of generative AI model. For an enterprise, fine-tuning these models to a specific use case or for internal enterprise use in a specified industry more data is required, often proprietary in nature. In the case of LLMs, one option available to enterprises is to train and develop their own LLM. However, this is an expensive and time-consuming process and requires access to high-quality datasets and computing power which most businesses simply do not have access[18] to.

However, not a lot of information is available about the details of the data which is used to train the more powerful generative AI models. This poses a privacy risk to individual users, but also a reputational, legal, or financial risk to downstream business or enterprise users, as a lack of transparency or accuracy in the original training data set may lead to harms of bias and discrimination in output, regurgitation of proprietary information, or false results.[19] Therefore, it is important to examine these risks in more detail and deliberate on governance strategies for their mitigation.

Before delving into organizational privacy risks and mitigation strategies, the next sub-section squarely identifies the potential areas of conflict that exist in the functioning of generative AI models against the requirements of data protection laws and regulations globally.

Ensuring Adherence to Data Protection Norms

Globally enacted data protection regulations have achieved consensus on some core data protection requirements which are common to all jurisdictions.

The technical, architectural, and functional dimensions of generative AI challenge some of these notions and lead to ambiguities around the applicability of well-established data protection requirements.

For this reason, data protection regulators and supervisory authorities across the world have, over the past few months, actively undertaken investigations and examined the way generative AI offerings may undermine the data protection requirements envisaged specifically under the GDPR.[20]

The paragraphs below present a high-level overview of three key data protection challenges and ambiguities that can be identified across jurisdictions; first, identifying lawful grounds for processing personal data in the context of generative AI systems, second, adhering to commonly agreed upon data protection principles, and third, responding to data subjects' rights requests.

Identifying Grounds for Processing of Personal Data

Under most data protection laws, including the Digital Personal Data Protection Act (DPDPA) 2023 in India[21], there are defined and limited grounds for processing of personal data which may be relied upon by a business. In the EU, article 6 of the GDPR provides for the 'lawfulness of processing,' wherein consent, performance of contract, compliance with the law, protection of vital interests of data subjects, public interest, and legitimate interests of the data controller/third parties are the valid basis which can be relied upon by an entity seeking to process personal data.[22]

However, in the context of development and deployment of generative AI, identifying the grounds of processing personal data may not be a straitjacket formula. Consent may be an appropriate basis to rely on where the developer of a system has a direct relationship with the individuals whose data is sought to be processed.[23] However, in the case of generative AI systems where training data sets are developed, for instance, through web scraping, seeking consent is not viable. Under the GDPR, the 'legitimate interest' may seem more flexible, however, it is not

always appropriate as it requires an entity to undertake a 'three-part test' to ensure that no harm is being caused to individuals whose data is being processed.[24]

Similarly, relying on contractual obligations may also not be feasible in all circumstances. For instance, in the case of an AI-chatbot, the Italian data protection authority held that performance of contractual obligations cannot be relied upon to process personal data of children, as they lack the legal capacity to enter a binding contract.[25]

Adhering to Data Protection Principles

Principles of collection limitation, data minimization, accuracy, storage or retention limitation, lawfulness and fairness, etc. are recognized in most data protection regulations. However, concerns have been observed about the capability of generative AI models to ensure adherence to these principles. For instance, adherence to the accuracy principle is often challenging. Inaccurate personal data may form part of the original training data set, leading to outputs which do not match factual truths. Additionally, the accuracy principle also requires that these AI models ensure that information is updated, which is again challenging as the training process may sometimes rely on information sourced through data scraping up to a certain point of time.[26] Even where the ability to undertake real-time web browsing is feasible, accuracy of outputs remain challenging. The tendency of large language models to sometimes 'hallucinate' and generate inaccurate content is well-documented.[27] While some research in undertaking specific types of processing indicates that process-level supervision, instead of outcome-level supervision, may improve the overall accuracy in the performance of LLMs.[28] The underlying issue of accuracy still remains pervasive.

Another core principle of data protection regulations is the notion of data minimization, which requires data controllers to only collect limited personal data which is necessary and relevant to fulfil stated purposes.[29] However, as explained in the preceding section, the development of generative AI models inherently relies on collection and processing of massive datasets, often containing personal data.

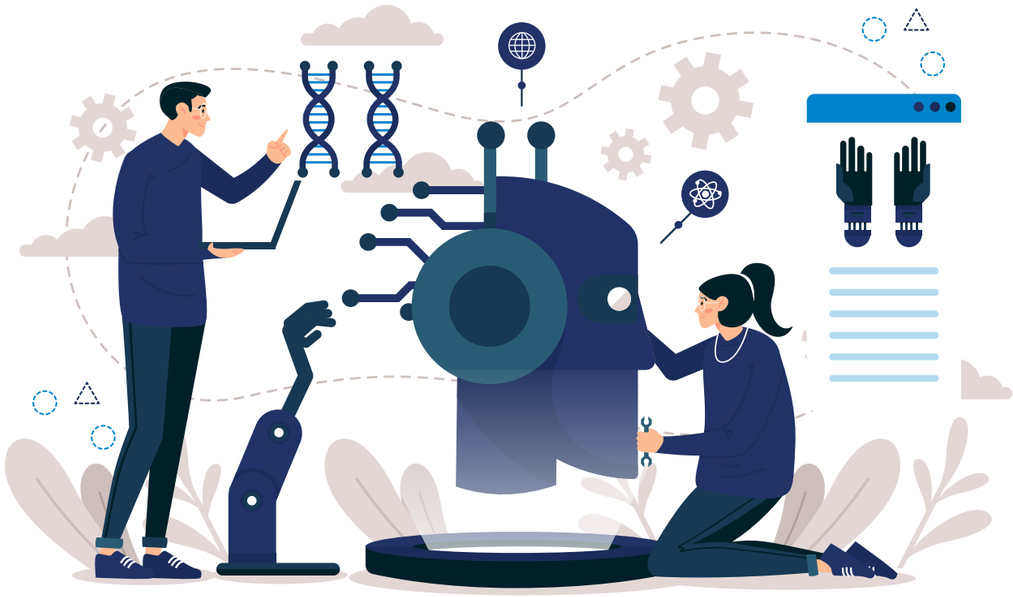
In the case of developers of general-purpose AI systems, not all use-cases can be reasonably foreseen at the instance of development of a model. This may make it difficult to adhere to the purpose limitation principle. For developers of generative AI models, it would therefore be useful to define and identify purposes in a narrow and precise manner, including detailed descriptions of the type of model developed, its technically feasible functionalities and downstream usages, conditions for use of the AI system, etc.[30]

Responding to Data Subjects' Rights Requests

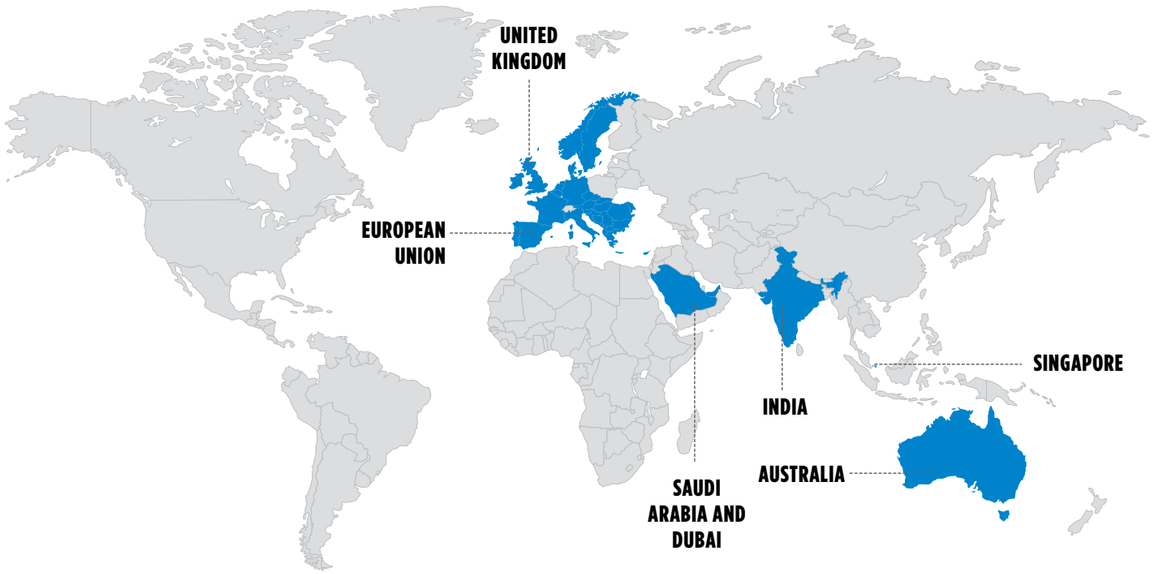
Data protection regulations globally recognize certain legally enforceable rights of data subjects. Under the GDPR, this refers to the right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, and right to object in the context of automated decision-making.[31] In India, under the DPDPA, the framework of data protection rights varies a bit, with the recognition of right to access and right to rectification along with the right to nominate and grievance redressal.[32]

However, unlike the GDPR, the right to object and the right to data portability do not find a mention in the Indian legal framework. Exercise of these rights in the context of generative AI models poses novel challenges. Specifically, the right to correction and erasure is challenging to implement due to the multiple sources of data during the training phase, the lack of attribution for these sources, and the embedded nature of personal data.[33]

Recent research highlights the challenges associated with exercising the right to be forgotten against large language models where proprietary training datasets are not publicly disclosed, as opposed to search engines where data is organized through indexing and can be easily accessed by individuals through querying.[34] However, some efforts have been made in this direction by leading generative AI providers to empower individuals to exercise their data protection rights and exercise greater control over the personal data which may be shared in prompts to a model. This includes enabling users to turn off chat history to ensure that user inputs are not used to train the language models and ensuring that data shared by enterprises using the API version of the service users' data is not used to train models by default.[35]



Global Data Protection Regimes and Generative AI Systems



Source: https://www.surveyofindia.gov.in/pages/world-map
Disclaimer: The World map used here is only for general illustration purpose.

While there are common foundational requirements of data protection regulations across jurisdictions, the proliferating use of generative AI has led to diverse approaches in how each regulator applies their national law in this context. This section aims to provide a jurisdictional analysis of how data protection laws, enforcement decisions, and guidance have been interpreted to adapt to the novel and complex functioning of generative AI.

INDIA

In India, the DPDPA 2023 has a distinctive framework where certain exemptions may impact processing of personal data for AI training directly. The scope of the Act does not extend to personal data which is publicly available where either the data principal has themselves made the data available or where the data was made publicly available by another person, pursuant to a legal obligation.[36] In practice, this provision may be applicable to scenarios where, for example, a social media user shares personal data and makes it publicly available and accessible.[37]

Further, the Act is also not applicable in its entirety where processing of personal data is necessary for research or statistical purposes.[38] This provision may also arguably impact the development and deployment of generative AI systems in India. However, this exception is limited to scenarios where no decision is being made about the data principal and subject to standards that will be prescribed by the Central Government.[39]

AUSTRALIA

In Australia, the Australian Privacy Principles (APPs) are the underlying foundation of the Privacy Act, 1988. The thirteen principles prescribe the broad standards and mandates that an organization should adhere to while processing personal information of individuals.[40] While the Privacy Act in Australia applies generally to all personal information, the Digital Platform Regulators Forum (DPRF) has expressed some privacy concerns emerging out of the functionalities of LLMs. The DPRF, which also includes the Office of the Australian Information Commissioner (OAIC) as a member, in its working paper on LLMs observed that the design of LLMs makes it difficult to ensure transparency in the handling of personal information.[41] Other privacy concerns highlighted include the heightened risk of data breaches, especially in the context of sensitive data, hallucinations in AI-generated output leading to misleading information about individuals, the reduced agency of individuals in scenarios where their personal data is used for training of generative AI models without their consent.[42]

At the intersection of privacy and online safety concerns, the e-Safety Commissioner's position statement on generative AI highlights that chatbots based on generative AI and other multimodal models may be capable of precise personalization of responses and outputs which may then consequently lead to personalized phishing or defrauding attempts to gain access to systems or personal information.[43] Parallelly, the commissioner opines that the potential of generative AI can also be leveraged to improve privacy standards and awareness. For instance, in the context of consent, it may be worth examining use cases of conversational AI being used to ensure nuanced and specific consent from individuals.[44]

EUROPEAN UNION

In the European Union, several national data protection authorities have independently initiated investigations into the functioning and impact of generative AI systems, as outlined in the previous section. Poland[45], Italy[46], Spain[47], and France[48] commenced independent investigations and examinations to determine the data protection impact of generative AI systems. In France, the CNIL has released its action plan for AI systems focusing its efforts on understanding the impact of AI technologies, including generative AI, on people, facilitating the development of systems which respect personal data, supporting innovation in the AI ecosystem, and developing mechanisms for audit and control.[49]

Parallelly, there have also been multilateral and collaborative efforts made towards understanding how data protection laws in the EU are applicable to generative AI, and what regulatory adaptations may be required in this context. An initiative of the European Data Protection Supervisor (EDPS), for instance, led to the 45th Global Privacy Assembly adopting a resolution on GenAI systems with global data protection supervisors as signatories.[50] The resolution recognizes and acknowledges the potential for data protection risks and harms that may arise out of generative AI systems, including collection of personal data from publicly accessible sources and in the context of automated decision making. The resolution re-emphasizes that current data protection regulations continue to apply. This includes requirements for a lawful basis for processing, purpose and use limitation prohibiting incompatible secondary use of personal data, adhering to principles of data minimization and accuracy, privacy

by design and default, transparency and accountability, and finally providing means for data subjects to exercise their rights.

To ensure coordination and information exchange amongst national data protection authorities, the European Data Protection Board also launched a dedicated taskforce to examine data protection related concerns arising from the use of ChatGPT.[51]

SAUDI ARABIA AND DUBAI

Over the last two years, jurisdictions in the Middle East have also shifted their regulatory focus towards addressing the privacy and security risks and opportunities emerging from the use of generative AI. In Saudi Arabi, the Saudi Data and AI Authority (SDAIA) has released a comprehensive guidance on the development and use of generative AI.[52] The guidance emphasises on the benefits and opportunities arising from Generative AI in the form of improved efficiencies, informed decision making, increase in quality of public service delivery, etc. The guidance provides insights on privacy considerations for both 'users' and 'developers' of generative AI systems, while recommending the incorporation of principles of privacy and security by design and undertaking continuous privacy impact and risk assessments.[53]

In Dubai, the Dubai International Financial Centre (DIFC), has recently amended Regulation 10 of the Data Protection Regulations, by introducing new mandates for processing of personal data by autonomous and semi-autonomous systems, including generative AI systems. The amended regulations impose obligations on 'deployers' and 'operators' of AI systems, where the deployer may be understood as a person, entity, or organization which is utilizing the output of an AI system or has authority over its functioning. An 'operator' may be understood as the provider of an AI system for the benefit of and at the direction of the deployer.[54]

SINGAPORE

In Singapore, the regulatory and governance focus in the context of artificial intelligence is on developing its digital economy by fostering a trusted ecosystem. The country's efforts in this direction are highlighted by the creation of the 'AI Verify' governance and testing framework which relies on eleven ethical principles, including data governance.[55] To facilitate innovation while ensuring that generative AI technologies are also safe and secure, the Infocomm Media Development Authority (IMDA) in Singapore has also announced a 'Generative AI Evaluation Sandbox' which aims to set out foundational and baseline recommendations for LLMs.[56]

More recently, in its proposed 'Model AI Governance Framework for Generative AI', the IMDA and AI Verify Foundation, highlight three focus areas for policymakers when considering the use of data in development of generative AI; ensuring trusted use of personal data, balancing copyright with data accessibility, and facilitating access to quality data.[57] On the use of personal data, some high-level suggestions and observations include exploring the use of privacy enhancing technologies to safeguard confidentiality and privacy.

UNITED KINGDOM

In the United Kingdom, the Information Commissioner (ICO) is actively undertaking a series of consultations on the intersection and applicability of existing data protection regulation on generative AI. The foremost issue undergoing consultation is centred around the lawful basis for scraping data, including personal data, to train generative AI models.[58] While outlining the various sources from which data may be obtained in the training and development process (publicly accessible datasets put together by third parties, directly undertaking web-scraping, etc.), the ICO emphasizes that data protection obligations continue to be applicable to developers.

In the policy position laid out in the consultation, a preliminary analysis indicates that five of the six lawful bases available under the UK GDPR would be inapplicable in the context of usage of web-scraped data for generative AI training, thereby limiting the available lawful basis to 'legitimate interest.' The key questions underlying the consultation hinge on determining the nature of safeguards and guardrails that developers of generative AI need to implement in order to satisfy the three-part requirement of a legitimate interest, necessity for processing, and balancing individual rights with interests of the developer.

The U.K. Parliament's report on large language models and generative AI reemphasizes the need for further clarification on how data protection laws apply to complex large language models, specifically with regard to the ability of individuals to exercise their data protection rights where their personal data has been used in the training process.[59]

Data Protection Implications in Enterprise Use of Generative AI

Enterprise adoption and use of generative AI may create additional risks and challenges in the context of data protection and accountability. Technical, organisational, and enterprise policy-level interventions can help mitigate some of these risks.

Affixing Accountability for Data Protection in the AI Supply Chain

Data protection regulations globally recognize a dichotomous relationship of entities as either data fiduciaries/controllers or as data processors. Some regulations, such as the GDPR, also recognize the relationship of 'joint data fiduciaries'. However, in the context of development and deployment of generative AI systems, these traditional conceptual notions may be challenged, and the role played by each entity in the supply chain may not be as clearly delineated.

As highlighted in previous sections, most of the data used during the training of generative AI models is obtained through web scraping. Developers of generative AI models may not be accurately classified as data controllers in this context because the personal data used in the original training dataset, its collection and the purposes for which it was collected, and the output were not determined by the developer of the AI model. However, it may also not be feasible to categorize them as data processors. Data processors are understood as entities which process personal data on behalf of the data controller/fiduciary[60], and such processing takes place based on a contractual relationship.[61] Both these elements are generally absent in this context, making it difficult to qualify developers as 'data processors'.

In the context of API use of generative AI models for downstream application by enterprises, such organizations may qualify as data fiduciaries/controllers, as they define the purposes for which personal data is processed. For instance, when using OpenAI services via the OpenAI API platform, an order processing relationship is established between the data controller as the client and OpenAI as the processor. For these purposes, OpenAI provides a Data Processing Agreement.[62] However, it is also conceivable that the developer of the language model and the enterprise user be treated as joint controllers. OpenAI, however, still does not provide a template for a contract pursuant to Art. 26 GDPR for establishing a joint controller relationship.

Affixing accountability and responsibility for data protection along the AI value chain can be difficult to navigate as the relationships between developers and deployers are complex. For instance, when a large language model is deployed by an enterprise, which then modifies it for its own internal use, it may be difficult to pin down which performance issues are to be attributed to the enterprise (i.e. the deployer) and which are to be attributed to the developer of the language model.[63] In the discussions and debates leading up to the formalization of the EU AI Act as well, it was observed that the value chain in the context of artificial intelligence makes it important to also consider what obligations should persist for relevant third-parties as the manner in which a model is deployed or used by a downstream user can impact the functioning of the model and therefore have an impact on outcomes of the model.[64]

From a regulatory standpoint, therefore, it becomes important to understand the roles of different stakeholders at different junctures of the AI technology stack, i.e., the application layer, the model layer, and the infrastructure layer.[65] Simultaneously, from an enterprise standpoint, it becomes important to determine each stakeholder's respective role and legal obligations, on a case-by-case basis, pertaining to data protection corresponding to their role. In the context of AI systems, specifically large language models, where the developer of an AI system is responsible for also creating the training data sets. They may be qualified as a controller/fiduciary, where two or more controllers jointly determine the means and purpose of processing they may qualify as joint controllers, and finally, where an organization develops an AI system on behalf a customer they may qualify as a processor.[66]

Ensuring Responsible Use of Foundation Models by Enterprises

In addition to understanding the roles and legal obligations on an enterprise when utilizing generative AI services, it is also important to emphasize on measures for facilitating responsible use of generative AI systems at an organizational level.

It is well-documented that large language models may be capable of systemic risks such as discriminatory outputs, misinformation, and dissemination of private, confidential, or sensitive information.[67] Privacy related risks may include leaking of personal information which formed a part of the original training data set or may also entail inferences made about an individual which are factually incorrect.[68]

It is, therefore, pertinent that organizations seeking to integrate generative AI into their business processes adopt measures to ensure responsible use of the technology, whether the use-cases are external facing, for example in the case of an AI-enabled customer chatbot, or for internal use, or use of generative AI as a part of decision-making processes. Some of the best practices that may be adopted by enterprises are highlighted below. These suggestions are based on guidance

Privacy related risks may include leaking of personal information which formed a part of the original training data set or may also entail inferences made about an individual which are factually incorrect.

issued and consultations undertaken in Australia[69], Canada[70] and Germany[71], as well as checklists released by think tanks[72] and industry stakeholders[73] on effective enterprise-level governance of generative AI.

- **Informed use of generative AI for defined purposes:** Generative AI systems should be adopted by enterprises after a thorough examination of the privacy practices and commitments of the developer of the AI system. Furthermore, to strengthen commitment to the principles of data minimization and purpose limitation under data protection regulations, it is recommended to outline and define the specific purposes for which a generative AI system may be deployed, avoiding prompts and uses which may lead to identification of otherwise de-identified or anonymized data. Purpose or function creep should be avoided by enterprises by limiting the use of the generative AI tools or systems to functions which are justifiable.

- **Ensure transparency about organizational use of generative AI:** Where generative AI systems are external-facing, for example in the case of customer chatbots, clear communication must be made to the relevant third-party about the use of such systems, how their personal data will be used, safeguards deployed by the enterprise, and the risks associated with an individual's interaction with such systems. In the case of chatbots specifically, providing customers with an option to 'opt-out' of their personal data being used for further training of the model would also strengthen individuals' agency over their personal data.

- **Establish accountability and explainability:** Prior to deploying generative AI tools at scale, privacy or data protection impact assessments could help gauge the impact on data protection compliance obligations, rights of individuals, and other privacy considerations. Both developer and deployer or user of a generative AI system must aim towards making the tool or system explainable by developing capabilities to be able to provide a structured and complete account of how the system works and the rationale behind any specific output produced.

- **Create mechanisms for human oversight:** Despite the rapid advancement in the capabilities of generative AI tools, there remains scope for generative AI tools and models to 'hallucinate' and produce outputs which are not representative of factual reality. Therefore, where these systems are being relied upon for any decision-making in an enterprise or being used otherwise in high-risk scenarios, enterprises must ensure that a human reviewer verifies the output or recommendations generated by the AI system for veracity.

- **Implement governance and auditing mechanisms with continuous evaluation:** Where feasible, concrete governance and auditing mechanisms should be instituted in enterprises to ensure that records are maintained of any privacy-related breaches or incidents and the same is also reported as feedback to the developers of the system or tool. Additionally, it is also important to evaluate any generative AI system or tool continuously and regularly for vulnerabilities and to identify potential use cases where data protection harm may occur. To this end, red teaming of AI systems, where viable, could be worth undertaking.

- **Undertake risk assessment for service providers/developers:** Where enterprises leverage pre-existing generative AI services and tools, they should undertake an assessment of the developer or service provider to assure that they will be able to comply with applicable data protection laws, respond to requests for deletion or rectification if required, implement appropriate and adequate technical and organizational measures to safeguard the data, etc.

Implementing Safeguards for Privacy Preservation

The previous section highlighted some of the organizational safeguards that can be implemented by an enterprise seeking to integrate existing generative AI systems in their business processes and workflows. However, a comprehensive strategy to mitigate privacy risks emerging from generative AI should also entail examining technical safeguards for privacy preservation, especially at the development stage of these models.

Since the launch of public-facing generative AI services, like chatbots or image-generation tools, there has been a steep rise in the actions of mala fide actors attempting to bypass inbuilt security and safety guardrails of these generative AI systems. One such intervention is popularly known as 'jailbreaking,' which in the context of generative AI, refers to deliberately designing prompts in a manner which is intended to override restrictions coded into AI programs to facilitate generation of illegal or harmful content.[74] Some research indicates that though popular text-to-text generative AI services may be resilient to direct prompts seeking to leak personal information, the safety features built into these services may not fully be capable of defending against multi-step jailbreaking prompts.[75]

Another type of closely related attack is the use of 'prompt injections' to direct the generative AI model to create harmful or potentially illegal output. This may be done through direct or indirect prompt injections, where the former entails prompts generated by the user of the generative AI system while the latter entails discreet, malicious prompts from a third-party source such as a website or a document being analyzed by the AI system.[76]

While developers of AI have been actively working fixing vulnerabilities to make generative AI systems more resistant to jailbreaking and prompt injection attacks, there may be a requirement to look beyond traditional red-teaming methods and deploy automated and advanced techniques to mitigate these attacks at scale.[77]

It is relevant to emphasize here that despite best efforts from developers, it may still be possible to extract personal data and sensitive information from large language models.[78] From an enterprise perspective, where open-source models are utilized or generative AI models are used through API-access and fine-tuned for customized use-cases in an enterprise, some fine-tuning interventions can themselves lead to divulging of personal data.[79]

Where training of generative AI entails learning patterns and features of personal data, such personal data invariably becomes a part of the AI model and

influences its outputs.[80] Therefore, it is important that personal data should only be included in training where there is a compelling need for the same in terms of the intended operational objectives of the generative AI system, ideally after undertaking a data protection impact assessment.[81]

From the perspective of a downstream enterprise user of a generative AI system, it is also important to carefully consider which categories of personal data is shared with the system. Once personal data has been shared with generative AI models, monitoring its storage and usage across various systems or its retraction may be an arduous task.[82]

Some key interventions that can be incorporated by enterprises in their AI governance strategies, to mitigate risks pertaining to unauthorized or unprecedented disclosure of personal data include:

i Ensuring accuracy of data for verifiable results,

ii Undertaking efforts to reduce harmful outputs through bias, explainability, robustness, and security assessments,

iii Adhering to honesty and transparency principles by respecting data provenance and using user-consented data where feasible,

iv Using zero-party or first-party data for training of AI models, and

v Designing processes with a human-in-the-loop to review outputs and automated decision-making.[83]

The use of data redaction and the contextually accurate synthetic data in the training process of generative AI models has also been proposed as a solution.[84] However, data redaction of unstructured data at scale is difficult at scale, while manually redacting data is slow and expensive, as well as often inaccurate. Further, where synthetic data used for training or fine-tuning is generated by another generative AI model, newer and more complex challenges emerge in terms of poor-quality outputs, what current research terms as 'Model Autophagy Disorder'.[85]

CONCLUSION

As the adoption of LLMs and GenAI continues to surge, it becomes increasingly apparent that these technologies offer immense potential for driving business innovation. However, alongside this promise, lurks the shadow of significant cybersecurity risks if not managed effectively. It is imperative to recognize that security is not an afterthought but rather the cornerstone of responsible AI development.

Enterprises must prioritize the implementation of proactive and comprehensive cybersecurity strategies to safeguard their digital assets and uphold stakeholder trust. This entails the establishment of robust governance frameworks, the adoption of zero-trust platforms, the reinforcement of enhanced controls, continual employee training, and strict adherence to regulatory compliance.

Collaboration among cybersecurity leaders, technology teams, and risk management experts is paramount for effectively navigating these evolving risks. Developers, too, bear a crucial responsibility in adopting responsible practices throughout the development lifecycle of LLMs, including purposeful design, transparency, and the integration of robust security measures.

At the intersection of data protection and generative AI, there are several regulatory grey areas that will require further research and innovation to resolve the inherent conflict of massive datasets used for training, data protection obligations under regulations, and privacy rights of individuals.

Across jurisdictions, there is largely a consensus that data protection regulations will continue to be applicable to generative AI technology, tools, and systems. However, there is ambiguity surrounding the precise manner in which data protection obligations can be operationalized, for instance, in the context of exercising rights of data subjects/data principals. The technical architecture of generative AI and its training and development process makes it difficult to provide complete assurance about the ability of a developer or user of such tools to exercise complete control over its outcomes, downstream uses, and the inferences drawn by the model upon deployment.

Regardless of regulatory ambiguity, it is apparent that while sustained efforts are continuously being directed towards improving the privacy protections of generative AI technologies, there remain gaps in the privacy assurance of these novel and complex tools. At an enterprise level, this translates into a requirement to invest in resources, training and awareness, and implementation of best practices to mitigate any risks arising from the use of generative AI. At the most fundamental level, enterprises should conduct impact assessments to understand privacy risks from the use of any generative AI technology and adhere to the fundamental data protection and ethical principles of accountability, fairness and transparency, explainability, and proportionality.

References

1. ‘Definition of Generative AI - Gartner Information Technology Glossary’, Gartner <https://www.gartner.com/en/information-technology/glossary/generative-ai>.
2. ‘The History of Artificial Intelligence - Science in the News’ <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>.
3. ‘Learn to Build an AI Strategy for Your Business’, Gartner <https://www.gartner.com/en/information-technology/topics/ai-strategy-for-business>.
4. Md Al-Amin and others, ‘History of Generative Artificial Intelligence (AI) Chatbots: Past, Present, and Future Development’ (arXiv, 2024) <https://doi.org/10.48550/arXiv.2402.05122>.
5. ‘Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026’, Gartner <https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026>.
6. ‘AI Governance Frameworks For Responsible AI | Gartner Peer Community’ <https://www.gartner.com/peer-community/oneminuteinsights/ai-governance-frameworks-responsible-ai-33q>.
7. ‘MITRE | ATLASTM’ <https://atlas.mitre.org/>.
8. Apostol Vassilev, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (Gaithersburg, MD: National Institute of Standards and Technology, 2024), p. NIST AI NIST AI 100-2e2023 <https://doi.org/10.6028/NIST.AI.100-2e2023>.
9. Generative AI and Authenticity | Centre for Trustworthy Technology <https://c4tt.org/wp-content/uploads/2024/01/Generative-AI-and-Authenticity.pdf/>.
10. ‘Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models’, ICO <https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/>.
11. ‘ChatGPT: Everything you need to know about OpenAI's GPT-4 tool’, BBC Science Focus <https://www.sciencefocus.com/future-technology/gpt-3/>.
12. GPT-4 Technical Report | OpenAI <https://cdn.openai.com/papers/gpt-4.pdf/>.
13. ibid
14. Common Crawl <https://commoncrawl.org/>.
15. LAION : Large-scale Artificial Intelligence Open Network <https://laion.ai/>.
16. ‘Your Personal Information is Probably Being Used to Train Generative AI Models’, SCIAM <https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/>.
17. ‘Joint statement on data scraping and the protection of privacy’, ICO <https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf>.
18. ‘How to Train Generative AI Using Your Company’s Data’, Harvard Business Review <https://hbr.org/2023/07/how-to-train-generative-ai-using-your-companys-data/>.
19. ‘What is generative AI?’, IBM <https://research.ibm.com/blog/what-is-generative-AI>.
20. ‘HOW DATA PROTECTION AUTHORITIES ARE DE FACTO REGULATING GENERATIVE AI’, FUTURE OF PRIVACY FORUM <https://fpf.org/blog/how-data-protection-authorities-are-de-facto-regulating-generative-ai/>.
21. Digital Personal Data Protection Act 2023, section 4, <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf/>.
22. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, Article 6, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679/>.
23. ‘How do we ensure lawfulness in AI?’, ICO <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-lawfulness-in-ai/>.
24. ibid
25. ‘Artificial intelligence: italian SA clamps down on ‘Replika’ chatbot’, Garante <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852506#english>.
26. Generative AI: The Data Protection Implications | CEDPO <https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf>.
27. ‘What are AI hallucinations?’, IBM <https://www.ibm.com/topics/ai-hallucinations> [accessed 6 March 2024]., ‘Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive’, Stanford HAI <https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive>, ‘Chatbots May “Hallucinate” More Often Than Many Realize’, NYTimes <https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html/>.
28. Improving mathematical reasoning with process supervision | OpenAI <https://openai.com/research/improving-mathematical-reasoning-with-process-supervision>.
29. ‘Principle (c): Data Minimisation’, ICO <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/>.
30. ‘Defining a purpose’, CNIL <https://www.cnil.fr/en/defining-purpose/>.
31. General Data Protection Regulation, 2016, Articles 15 to 21, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504>.
32. Digital Personal Data Protection Act 2023, <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf>.
33. supra 26.
34. Dawen Zhang and others, Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions (arXiv, 2023) <https://arxiv.org/pdf/2307.03941.pdf/>.
35. ‘New ways to manage your data in ChatGPT’, OpenAI <https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt>.
36. Digital Personal Data Protection Act 2023, Section 3(c)(ii), <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf>.
37. ‘THE DIGITAL PERSONAL DATA PROTECTION ACT OF INDIA, EXPLAINED’, FUTURE OF PRIVACY FORUM <https://fpf.org/blog/the-digital-personal-data-protection-act-of-india-explained/>.
38. Digital Personal Data Protection Act 2023, Section 17(2)(b), <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf/>.

39. Digital Personal Data Protection Act 2023, Section 40(2)(q), <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf/>.
40. Australian Privacy Principles quick reference | Office of the Australian Information Commissioner <https://www.oaic.gov.au/privacy/australian-privacy-principles/australian-privacy-principles-quick-reference>.
41. Working Paper 2: Examination of technology – Large Language Models | Digital Platforms Regulators Forum <https://dp-reg.gov.au/publications/working-paper-2-examination-technology-large-language-models#potential-impacts-in-regulated-areas>.
42. ibid
43. Tech Trends Position Statement Generative AI | eSafety Commissioner <https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf>.
44. ibid
45. 'The technology must be compliant', UODO <https://uodo.gov.pl/pl/138/2823>.
46. 'Artificial intelligence: stop to ChatGPT by the Italian SA', Garante <https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english>.
47. 'The AEPD initiates ex officio investigation proceedings against OpenAI, owner of ChatGPT', AEPD <https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/aepd-inicia-de-oficio-actuaciones-de-investigacion-a-openai>.
48. 'French privacy watchdog investigating complaints about ChatGPT', REUTERS <https://www.reuters.com/technology/french-privacy-watchdog-investigating-complaints-about-chatgpt-2023-04-11/>.
49. 'Artificial Intelligence: the action of the CNIL', CNIL <https://www.cnil.fr/en/artificial-intelligence-action-plan-cnil/>.
50. Resolution on Generative Artificial Intelligence Systems | Global Privacy Assembly <https://www.edps.europa.eu/system/files/2023-10/edps-gpa-resolution-on-generative-ai-systems_en.pdf>.
51. 'EDPB resolves dispute on transfers by Meta and creates task force on Chat GPT', EDPB <https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en/>.
52. GENERATIVE ARTIFICIAL INTELLIGENCE GUIDELINES | SDAIA <https://sdaia.gov.sa/en/SDAIA/about/Files/GenerativeAIPublicEN.pdf/>.
53. ibid
54. FAQs: REGULATION 10 ON PERSONAL DATA PROCESSED THROUGH AUTONOMOUS AND SEMI-AUTONOMOUS SYSTEMS | DIFC <https://edge.sitecorecloud.io/dubaiintern0078-difcexperie96c5-production-3253/media/project/difcexperiences/difc/difcwebsite/documents/data-protection-pages/guidance-and-handbooks/lawful-processing/difc-dp-gl-24-rev-01-regulation-10-faqs.pdf>.
55. 'Singapore's Approach to AI Governance', PDPC <https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework>.
56. 'First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA', INFOCOMM MEDIA DEVELOPMENT AUTHORITY <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox/>.
57. PROPOSED MODEL AI GOVERNANCE FRAMEWORK FORGENERATIVE AI | AI VERIFY FOUNDATION <https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf/>.
58. 'Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models', ICO <https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/>.
59. 'Large Language Models and generative AI', Communications and Digital Committee U.K. Parliament <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm>.
60. Digital Personal Data Protection Act 2023, section 2(k), <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf/>.
61. Digital Personal Data Protection Act 2023, section 8(2), <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf/>.
62. Data Processing Addendum | OpenAI <https://openai.com/policies/data-processing-addendum>
63. 'A Policymaker's Guide to Foundation Models', IBM <https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models>.
64. Cornelia Kutterer, 'Regulating Foundation Models in the AI Act: From "High" to "Systemic" Risk', <https://ai-regulation.com/wp-content/uploads/2024/01/C-Kutterer-Regulating-Foundation-Models-in-the-AI.pdf>.
65. Governing AI: A Blueprint for the Future | Microsoft <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw>.
66. 'Determining the legal qualification of AI system providers', CNIL <https://www.cnil.fr/en/determining-legal-qualification-ai-system-providers>.
67. Laura Weidinger and others, Taxonomy of Risks posed by Language Models (FAccT, 2022) <https://dl.acm.org/doi/pdf/10.1145/3531146.3533088>.
68. ibid
69. 'Interim guidance on government use of public generative AI tools - November 2023', Australian Government Digital Transformation Agency <https://architecture.digital.gov.au/guidance-generative-ai/>.
70. 'Principles for responsible, trustworthy and privacy-protective generative AI technologies', Office of the Privacy Commissioner of Canada <https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/>.
71. 'Checklist for using LLM-based chatbots', Datenschutz Hamburg <https://datenschutz-hamburg.de/news/checkliste-zum-einsatz-llm-basierter-chatbots>.
72. 'FPF releases generative ai internal policy checklist to guide development of policies to promote responsible employee use of generative ai tools', future of privacy forum <https://fpf.org/blog/fpf-releases-generative-ai-internal-policy-checklist-to-guide-development-of-policies-to-promote-responsible-employee-use-of-generative-ai-tools/>.
73. 'How do we best govern AI?', Microsoft <https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai/>, 'The AIdea of India: Generative AI's potential to accelerate India's digital transformation', EY <https://www.ey.com/en_in/ai/generative-ai-india-report>.
74. supra 26.
75. Haoran Li and others, Multi-step jailbreaking Privacy Attacks on ChatGPT (arXiv, 2023) <https://arxiv.org/pdf/2304.05197.pdf/>.

76. ‘Generative AI’s Biggest Security Flaw Is Not Easy to Fix’, WIRED <https://www.wired.co.uk/article/generative-ai-prompt-injection-hacking>.

77. Supra 26.

78. Milad Nasr and others, Scalable Extraction of Training Data from (Production) Language Models (arXiv, 2023) <https://arxiv.org/abs/2311.17035>.

79. Xiaoyi Chen, The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks (arXiv, 2023) <https://arxiv.org/pdf/2310.15469.pdf>.

80. supra 26.

81. Resolution on Generative Artificial Intelligence Systems | Global Privacy Assembly <https://www.edps.europa.eu/system/files/2023-10/edps-gpa-resolution-on-generative-ai-systems_en.pdf>.

82. supra 26.

83. ‘Managing the Risks of Generative AI’, Harvard Business Review <https://hbr.org/2023/06/managing-the-risks-of-generative-ai/>.

84. ‘Addressing Privacy and GDPR in ChatGPT and Large Language Models’, PrivateAI <https://www.private-ai.com/2023/01/18/addressing-privacy-and-the-gdpr-in-chatgpt-and-large-language-models/>.

85. Sina Alemohammad and others, Self consuming Generative Models Go MAD (arXiv, 2023) <https://arxiv.org/pdf/2307.01850.pdf>.

Authors:

Neha Mishra
Associate, Technical Research, DSCI

Shivangi Malhotra
Associate, Privacy & Policy, DSCI

Contributor:

Charu Sharma
Assistant Manager
Communications & Marketing, DSCI

Data Security Council of India (DSCI) is a premier industry body on data protection in India, setup by nasscom, committed to making the cyberspace safe, secure and trusted by establishing best practices, standards and initiatives in cybersecurity and privacy. DSCI brings together governments and their agencies, industry sectors including ITBPM, BFSI, telecom, industry associations, data protection authorities and think-tanks for policy advocacy, thought leadership, capacity building and outreach initiatives. For more info, please visit www.dsci.in

DATA SECURITY COUNCIL OF INDIA

Nasscom Campus, Fourth Floor, Plot. No. 7-10, Sector 126, Noida, UP - 201303

+91-120-4990253 | info@dsci.in | www.dsci.in

DSCI_Connect dsci.connect dsci.connect data-security-council-of-india dscivideo