Gen AI Governance

* **Discriminative AI**  
  Broadly speaking, discriminative AI classifies data but is not able to generate it. Important applications of discriminative AI include sentiment analysis, named-entity recognition, image classification, and optical character recognition. A common characteristic of discriminative models is that their outputs are limited to a predetermined and finite set of target classes, though this is not a hard requirement.

There are a large number of commonly-used discriminative model classes that produce efficient, high-speed classifiers on fixed-size inputs. These include logistic regression, k-nearest neighbors, support vector machines, and gradient-boosted decision trees. Neural architectures such as convolutional neural networks (CNN) and long short-term memory (LSTM) units are often used to build reasonably-sized discriminative models for very long and varying-length inputs. For very large models, transformers—the neural component underlying the most recent advancements in AI—continue to gain popularity.

* **Generative AI**  
  Generative AI is capable of generating new content resembling the content on which it is trained. Popular Generative AI techniques include:
  + Generative Adversarial Networks (GANs)
  + Diffusion models
  + Autoregressive models

GANs are machine-learning techniques that consist of two neural networks, a generator and a discriminator. The generator generates data by shaping random noise fed to it into a target format (typically for images). On its own, it cannot assess the quality of its output. This is where a separate model, termed a discriminator, comes in.

The discriminator aims to differentiate between real data and fake data generated by the generator. The two are trained simultaneously, with the discriminator trained to differentiate real and generator data, and the generator trained to confuse the discriminator by making increasingly realistic data. As training progresses, each model becomes increasingly better at its task, resulting in the generator being able to create realistic-looking content.

Despite their unprecedented success, a persistent challenge with GANs is training them. For instance, GANs can undergo model collapse in training, in which the generator only learns to generate a small variety of samples sufficient to confuse the discriminator but not sufficient to be useful. A recent successor to GANs, having a much improved training regime, are diffusion models. In essence, diffusion models are trained to recover training data from noisy-fied versions of it. After training, diffusion may ideate entirely new images from a pure noise input. Many popular image-generation services are built on diffusion models.

Autoregressive models are the oldest of the three generative approaches described in this section, having their roots in the field of statistics rather than machine learning. Autoregressive models generate sequences of data by modeling the probability of the next element in a sequence conditioned on the prior elements. The next element is then randomly selected from this distribution, using a “temperature” parameter can nudge the results to be more deterministic or more random, and the process is repeated (much like writing). Popular neural network components for autoregressive models once again include LSTMs and transformers, the latter of which underlies the most impressive generative AI to date.

Until recently, autoregressive text models were challenging to use because they could only complete a sequence fed to it. To improve their utility, an additional alignment stage is performed. In alignment, the autoregressive model is additionally trained to prefer certain input-output pairs to others based on human feedback. In the case of generative large language models, alignment has successfully taught models how to respond to questions and commands. Alignment is typically performed using techniques from reinforcement learning.

Recent gains in generative AI have resulted from training very large generative models (100B+ of parameters) on substantial amounts of data (10TB+) and aligned using human feedback. Generative LLMs, in particular, now have sufficient knowledge and response accuracy to provide unprecedented zero-shot/few-shot accuracy, enough to supplant the use of dedicated discriminative classifiers for specific tasks.

**Significance of Artificial Intelligence**

The rapid development of AI systems and models, particularly since the launch of ChatGPT in November 2022, has profoundly energized the business landscape in an unprecedented manner. Generative AI is revolutionizing industries with significant advances in productivity and new capabilities.

Following are some examples of the unprecedented opportunities presented by AI to the business world:

* Automation for Efficiency: AI can automate repetitive tasks, leading to increased productivity and operational efficiency.
* Data-Driven Insights: AI has the capability to extract valuable insights from large datasets, providing businesses with a competitive edge through data-driven decision-making.
* Creative Problem Solving: AI can generate innovative solutions and ideas, even when provided with ambiguous or incomplete instructions, enhancing problem-solving and creativity.
* Content Creation: AI can produce high-quality content swiftly and on a large scale, benefiting content marketing, advertising, and customer engagement.
* Autonomous Decision-Making: AI enables levels of autonomous decision-making that were not possible with prior generations of AI.

**Generative AI and Its Value**

* Generative AI is a category of AI that excels at creating new content after learning patterns in real-world data. When provided with inputs or prompts, various generative AI models can generate diverse types of content. Here are some examples:
* Text Generation Models: Text generation models that have been aligned—typically through Reinforcement Learning from Human Feedback—include OpenAI ChatGPT, Google PaLM 2, and Meta LLaMA-2-Chat. These models are capable of unprecedented (albeit imperfect) capabilities in instruction following, which has led to their adoption across many industries. Particularly surprising are their abilities to perform zero-shot and few-shot learning, language translation, programming, and fluently generating meaningful content across a vast number of domains.
* Text-to-Image Models: Certain generative AI models, such as those underlying Stable Diffusion, Midjourney, and DALL-E, can produce, extend, or refine images from prompts.
* Text-to-Video Generation: Other models like Meta’s Make-A-Video can generate videos from prompts as well.
* AI models with generative capabilities, e.g., ChatGPT, DALL-E, etc., are also referred to by regulators as ‘general-purpose AI’ or ‘foundation models’. These AI models are trained on large sets of unlabelled data that can be used for different tasks with minimal fine-tuning.

Two key technologies underlying the generative AI revolution are (a) transformers, and (b) diffusion.

Transformers are typically used in text data but can be used for images and audio. They are the basis for all modern Large-Language Models (LLMs) because they allow neural networks to learn patterns in very large volumes of (text) training data. The result is the amazing capabilities observed in text generation models.

Diffusion models have overtaken Generative Adversarial Networks (GANs) as the neural models of choice for image generation. Unlike the error-prone image generation process of GANs, the “simplified” image generation process of diffusion models works by iteratively constructing an image through a gradual denoising process. The result is a myriad of new AI-based tools for generating and even editing images with useful outcomes.

According to McKinsey, just generative AI has the potential to contribute between $2.6 trillion and $4.4 trillion to annual business revenues. More than 75% of this value is expected to come from the integration of generative AI into customer operations, marketing and sales, software engineering, and research and development activities.

**Generative AI’s Need for Data**

Data plays a central role in the development of generative AI models, particularly Large Language Models (LLMs). These models rely on vast quantities of data for training and refinement. For example, OpenAI’s ChatGPT was trained on an extensive dataset comprising over 45 terabytes of text data collected from the internet, including digitized books and Wikipedia entries. However, the extensive need for data collection in generative AI can raise significant concerns, including the inadvertent collection and use of personal data without the consent of individuals. Google AI researchers have also acknowledged that these datasets, often large and sourced from various places, may contain sensitive personal information, even if derived from publicly available data.

Let’s explore the common sources of data collection employed by generative AI developers:

**Publicly-Accessible Data**

The majority of training data for generative AI comes from publicly accessible data sets. Web scraping is the most common method used to collect data. It involves extracting large volumes of information from publicly accessible web pages. This data is then utilized for training purposes, or may be repurposed for sale or made freely available to other AI developers.

Data obtained through web scraping often includes personal information shared by users on social media platforms like Facebook, Twitter, LinkedIn, Venmo, and other websites. While individuals may post personal information on such platforms for various reasons, such as connecting with potential employers or making new friends, they typically do not intend for their personal data to be used for training generative AI models.

**User Data**

Data shared by users with generative AI applications, such as chatbots, may be stored and used for training without the knowledge or consent of the data subjects. For example, users interacting with chatbots providing healthcare advice, therapy, financial services, and other services might divulge sensitive personal information. While such chatbots may provide terms of service mentioning that user data may be used to “develop and improve the service,” critics argue that generative AI models should seek affirmative consent from users or provide clear disclosures about the collection, usage, and retention of user data.

Considering their transformative potential, many organizations have also embedded generative AI models into their products or services to enhance their offerings. Such integration, in some cases, can also serve as a source of data, including the personal data of consumers, for the training and fine-tuning of these models.

**Significant Risks Associated with the AI**

While the business world increasingly recognizes the immense and unprecedented value brought about by the advancement of AI systems and models, there is also a growing global concern regarding the immediate dangers and risks associated with the unregulated progress of this technology.

The very qualities that make AI systems and models, such as LLM models, appealing technological innovations also render them potentially the riskiest technologies if not developed and implemented with careful consideration.

In particular, the current capabilities of AI models to learn patterns in vast quantities of data and make their insights available through natural language interfaces have real potential for the following abuses:

* Unauthorized mass surveillance of individuals and societies.
* Unexpected and unintentional breaches of individuals’ personal information.
* Manipulation of personal data on a massive scale for various purposes.
* Generation of believable and manipulative deep fakes of individuals.
* Amplifying while masking the influences of cultural biases, racism, and prejudices in legal and socially significant outcomes.
* Violation of data protection principles of purpose limitation, storage limitation, and data minimization.
* Discrimination against specific groups of individuals and societal bias.
* Disinformation and presenting factually inaccurate information.
* Intellectual property and copyright infringements.

The risks posed by the rapid advancement of AI systems and models have become so pronounced that, in an unprecedented move in March 2023, 30,000 individuals, including some of the world’s leading technologists and technology business leaders, signed a letter urging global governments and regulators to intervene unless AI developers agreed to voluntarily halt or slow down the development of AI technology for a period of six months.

**Overview: Introduction to AI Governance**

As we enter an era heavily influenced by generative AI technologies, the governance of artificial intelligence (AI) becomes an increasingly vital priority for businesses that want to enable the safe use of data and AI while meeting legal and ethical requirements. Globally, policymakers are perking up, paying attention, and taking action. In October 2023, the “safe, secure, and trustworthy” use of artificial intelligence warranted an executive order from the Biden-Harris administration in the US, an issuance that followed closely on the heels of the EU’s AI Act, the world’s first comprehensive AI law on the books. Other countries, like China, the UK, and Canada—and even a number of US states—have drawn their own lines in the sand, either proposing or enacting legislation that highlights the importance of safety, security, and transparency in AI. Recently, we have seen that regulatory authorities and courts are actively taking enforcement actions against AI systems; therefore, AI governance is becoming important for organizations.

AI governance is the way to this early adoption mindset—and is essential for enterprise leaders who are integrating AI services into their businesses. It addresses the challenge of managing compliance, safety, and security for the entire AI lifecycle—from creation to deployment. Effective AI governance provides control and oversight, ensuring that businesses develop and manage their AI services responsibly, ethically, and in compliance with both internal policies and external regulations in a documented, efficient, and demonstrable manner. This is crucial for businesses to navigate the complex landscape of AI technology while maintaining trust and accountability.

**AI Governance: Definition and Purpose**

AI governance refers to the imposition of frameworks, rules, standards, legal requirements, policies, and best practices that govern, manage, and monitor the use of artificial intelligence (AI). It involves directing, managing, and monitoring AI activities to meet legal and ethical requirements. On the ethical front, businesses are focused on ensuring a high level of transparency, safety, and security in their AI models to build and maintain customer trust. On the legal front, enterprises must conform to legal requirements and satisfy regulators—or risk substantial financial penalties and damaged brand reputation.

Hardly a day passes without news of a new technological or legal development in generative AI appearing in the media—and they all have implications for the way companies do business. This frenzy is justified, as McKinsey research estimates that generative AI could contribute between $2.6 trillion and $4.4 trillion in annual value going forward. However, to realize this potential, organizations must implement AI in a way that is transparent, secure, and trustworthy. In fact, Gartner suggests that organizations that successfully operationalize secure and trustworthy AI could see a 50% increase in their AI adoption and attainment of business goals.

The key drivers of AI governance in enterprises include innovation, efficiency, compliance, and trust—all crucial elements for successfully integrating AI into business operations.

* **Innovation:** AI governance provides a structured, yet flexible, framework that encourages responsible innovation. It allows for the development of advanced AI solutions within ethical boundaries, driving business growth and competitive advantage.
* **Efficiency:** By standardizing and optimizing AI development and deployment processes, AI governance enables enterprises to bring AI products to market faster while reducing costs.
* **Compliance:** AI governance aligns AI solutions and decision-making with industry regulations and global legal standards. This ensures that AI practices meet legal requirements, reducing legal risks for the business and furthering its regulatory compliance.
* **Trust:** AI governance focuses on building trustworthy and transparent AI systems—a practice that is crucial for maintaining customer rights and satisfaction while also protecting the organization’s brand value. Trustworthy AI enhances customer confidence and loyalty while reducing the risk of regulatory action.

**Gartner’s AI TRiSM: Tackling Trust, Risk, and Security in AI Models**

Recognizing the complexities of AI governance, industry experts like Gartner have developed specific frameworks to guide organizations in this area. Per Gartner, AI Trust, Risk, and Security Management (AI TRiSM) is a structured approach that will revolutionize businesses in the coming years. This framework, focusing on risk mitigation and alignment with data privacy laws in the use of AI, comprises four pillars: Explainability and model monitoring, Model operations, AI application security, and Model privacy.

**Explainability/Model Monitoring**

Model monitoring and explainability are crucial components in ensuring transparency and reliability in AI systems. They aim to provide clear explanations for the decisions or predictions made by AI models, facilitating a deeper understanding of their functioning. Regular monitoring of these models helps to verify their performance, identify potential biases, and comprehend their strengths and weaknesses. By elucidating details and reasons tailored to specific audiences, such as stakeholders or end-users, these practices enhance trust in AI systems and enable informed decision-making based on the likely behavior of the models. This also helps organizations fulfill the legal requirement of ensuring user transparency, i.e., informing users that they are interacting with an AI system, the logic of the AI system, and the rights individuals have with the help of privacy notices, instructions of use, and other similar mechanisms.

**Model Operations**

Model operations involve developing processes and systems for managing AI models throughout their lifecycle, from development and deployment to maintenance. Maintaining the underlying infrastructure and environment, such as cloud resources, is also a part of ModelOps to ensure that the models run optimally. This involves AI system governance and AI system classification.

**AI Application Security**

Since AI models often deal with individuals’ data that can be personally identifiable data or sensitive data, and any security breaches could have serious consequences, application security is essential. AI security keeps models secure and protected against cyber threats. So, organizations can use TRiSM’s framework to develop security protocols and **measures for safeguarding against unauthorized access or tampering.**

**Model Privacy**

Privacy ensures the protection of data used to train or test AI models. AI TRiSM helps businesses develop policies and procedures to collect, store, and use data in a way that respects individuals’ privacy rights. This is becoming important in industries such as healthcare, where sensitive patient data is processed using diversified AI models. Model privacy refers to managing data flows as per privacy laws, such as fulfilling data purpose limitation, storage limitation, data minimization, and other data protection principles.

AI TRiSM is an approach that is supposed to enhance AI models’ reliability, trustworthiness, security, and privacy. By using AI models more securely and safely, businesses can achieve improved goals, support various business strategies, and protect and grow their brands.

Let’s unpack some of the challenges or blind spots that lead to unregulated or uncontrolled AI.

Consider the issue of visibility into AI systems. When organizations lack clear insight into the deployment and operation of AI models, a phenomenon known as “shadow AI” can emerge. These unmonitored and unsanctioned models pose significant risks to security, ethics, and compliance, as they operate without proper oversight. This lack of visibility can lead to the perpetuation of biases, discrimination, and even malicious use of AI technologies.

Which models should you sanction and which ones should you block, is dependent on various risk parameters you would need to know at all times. Lack of awareness around the model risks may lead to issues like malicious use, toxicity, hallucinatory responses, bias, and discrimination.

The non-transparency surrounding the data used in AI models compounds these challenges. Lack of clarity regarding which data is being used in which AI model and which AI pipelines for training, tuning or inference may raise concerns about entitlements and the potential leakage of sensitive data. This lack of transparency not only undermines security but also raises concerns regarding compliance with data protection regulations.

Moreover, security controls for prompts, agents, and assistants powered by AI present a significant challenge. It’s crucial to understand how the data generated by these models is being utilized – whether it’s being shared in a Slack channel, integrated into a website as a chatbot, disseminated through an API, or embedded in an app. Moreover, these agents, while serving as channels for legitimate queries, also become potential pathways for new types of attacks on AI systems.

Finally, ensuring compliance with evolving industry standards and regulations adds another layer of complexity. From the NIST AI Risk Management Framework to the laws such as the EU AI Act and many other regulations in countries like Canada, China, Brazil, and Singapore, organizations must navigate a complex regulatory landscape to ensure their AI practices align with legal and ethical requirements.

Addressing these challenges requires a multifaceted 5-step approach that encompasses enhanced visibility into AI systems, comprehensive risk assessment, transparent data practices, robust security controls, and diligent compliance with regulatory frameworks.

**Step 1. Discover and catalog AI models**

This step aims to give businesses a complete and comprehensive overview of their AI usage by identifying and recording details of all AI models used in public clouds, private environments, and third-party apps. It covers the models’ purposes, training data, architecture, inputs, outputs, and interactions, including undocumented or unsanctioned AI models. Creating a centralized catalog of this information enhances transparency, governance, and the effective use of AI, supporting better decisions and risk management. It’s essential for revealing the full range of AI applications and breaking down operational silos within the organization.

**Step 2. Assess risks and classify AI Models**

This step allows businesses to assess the risks of their AI systems at pre-development and development stage and implement risk mitigation steps. This step also involves leveraging model cards that offer predefined risk evaluations for AI models, including a model’s description, intended use, limitations, and ethical considerations. These risk ratings provide comprehensive details, covering aspects such as toxicity, maliciousness, bias, copyright considerations, hallucination risks, and even model efficiency in terms of energy consumption and inference runtime. Based on these ratings, you can decide which models to sanction for deployment and use, which models to block, and which ones need additional guardrails before consumption.

**Step 3. Map and monitor data + AI flows**

Data flows into the AI systems for training, tuning and inference and data flows out of AI systems as the output. This step allows businesses to uncover full context around their AI models and AI systems i.e. map AI models and systems to associated data sources and systems, data processing, SaaS applications, potential risks, and compliance obligations. This comprehensive mapping enables privacy, compliance, security and data teams to identify dependencies, pinpoint potential points of failure, and ensure that AI governance is proactive rather than reactive.

Up to step three involves different levels of visibility into data and AI. Now, you need to implement guardrails to ensure safe data and AI usage.

**Step 4. Implement data + AI controls**

This step allows the establishment of strict controls for the security and confidentiality of data that is both put into and generated from AI models. Such controls include data security and privacy controls mandated by security frameworks and privacy laws respectively. For example, redaction or anonymization techniques may be applied in order to remove identifiable values from datasets. It ensures the safe ingestion of data into AI models, aligning with enterprise data policies and user entitlements. If sensitive data finds its way into LLM models, securing it becomes extremely difficult. Similarly, if enterprise data is converted into vector forms, securing it becomes more challenging.

On the data generation and output side, safeguarding AI interactions requires caution against external attacks, malicious internal use, and misconfigurations. To ensure secure conversations with AI assistants, bots, and agents, LLM firewalls should be deployed to filter harmful prompts, retrievals, and responses. These firewalls should be able to defend against various vulnerabilities highlighted in the OWASP Top 10 for LLMs, and in the NIST AI RMF frameworks, including prompt injection attacks and data exfiltration attacks.

**Step 5. Comply with regulations**

Businesses using AI systems must comply with AI-specific regulations and standards as well as data privacy obligations that relate to the use of AI. To streamline this demanding compliance process, businesses can leverage comprehensive compliance automation tailored to AI. Such a system offers a wide-ranging catalog of global AI regulations and frameworks, including the NIST AI RMF and the EU AI Act, among others. It facilitates the creation of distinct AI projects within its framework, enabling users to identify and apply the necessary controls for each project. This process includes both automated checks and assessments that require input from stakeholders, providing a holistic approach to ensuring compliance.

Enterprises that successfully carry out these five steps, will achieve –

* Full transparency into their sanctioned and unsanctioned AI systems,
* Clear visibility of their AI risks,
* Mapping of AI and data,
* Strong automated AI+Data controls, and
* Compliance with global AI regulations.