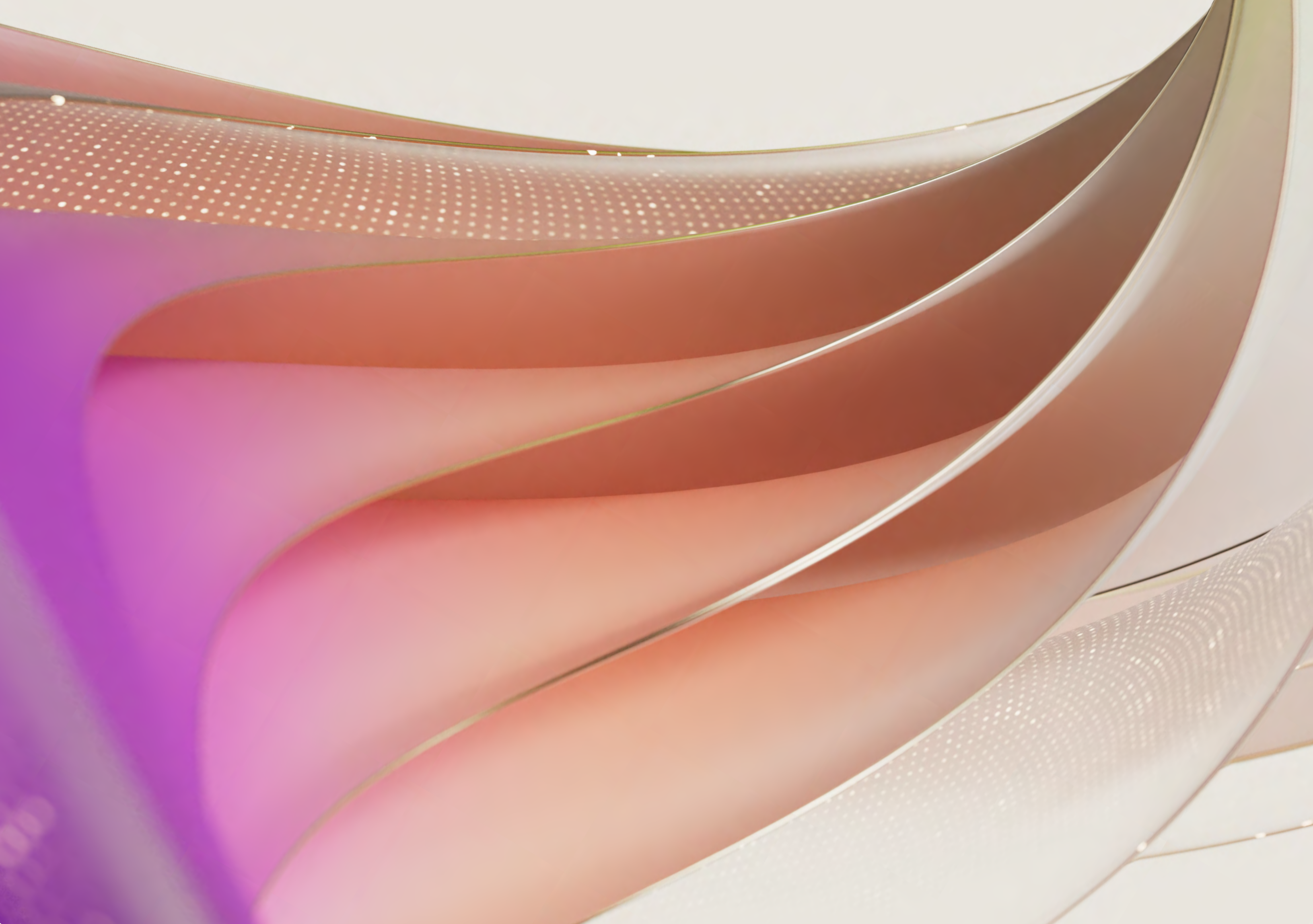
Microsoft

An Introduction to Generative AI and Safety

Concepts and definitions to gain a deeper understanding of generative AI



Introduction

Generative AI is an innovative and transformative technology that requires users to have deep understanding for success. Given the myriad of generative AI terms and concepts, this can be an intimidating undertaking without the right resources. Microsoft created this generative AI guide to give you clarity on this technology.

Use this guide to learn AI terminology, gain a deeper understanding of generative AI challenges, learn strategies to guide your organization in overcoming the biggest risks, and harness the true potential of AI. You'll find pertinent information in the following areas:

Definitions and applications

In this section, you'll learn AI terminology and understand the differences between predictive AI and generative AI so you can weigh the benefits of introducing generative AI into your own organization:

- Predictive AI
- Generative AI
- Machine learning
- Large language models
- Small language models
- Mechanics of generative AI models
- Generative AI applications

Predictive AI

Predictive AI can support problem solving and decision-making. To do this, it relies on symbolic reasoning, logic, and expert knowledge representation to emulate human reasoning. Models are built for specific purposes and usually work on large datasets. Techniques such as "knowledge representation languages," "rule-based reasoning," and "planning algorithms" are central to predictive AI. The advent of machine learning and deep learning techniques has shifted the focus of AI research toward learning from data. Popular use cases include IT process automation, threat detection, and business analytics.

Generative AI

Generative AI can create new content from simple prompts and context. These systems are trained on large volumes of data and then used as general purpose. They are trained to understand and respond to tasks like answering questions and writing essays. You can prompt generative AI to create text, code, images, videos, music, speech, game backgrounds, and more. Potential use cases include code generation, content creation, and product design. Gartner predicts that generative AI will directly impact the pharmaceutical, manufacturing, media, architecture, interior design, engineering, automotive, aerospace, defense, medical, electronics, and energy industries by augmenting core processes with AI models.

The differences between predictive AI and generative AI

While both predictive AI and generative AI are exciting technologies that can transform how organizations operate, there are several key differences between the two.

	Predictive AI	Generative AI
What it is	Common and usually integrated into the technology you already use	Next generation of AI in which models are trained to generate new original content based on natural language input
How it works	Follows set rules and completes a specific task proficiently, not creating anything new	Takes input from natural language, photos, or text to create new content
Popular examples	Voice assistants like Siri and Alexa and services that make recommendations like Netflix	Microsoft Copilot and ChatGPT
Opportunities	Assists humans rather than replaces them by answering queries much faster than a person could	Augments human skills rather than replaces them, helping humans offload rote tasks and derive more value from their time and energy
Outcomes	Outcomes usually accurate because of training methods and pattern recognition/repeatability	Outcomes vary, even if input is the same information

Machine learning (ML)

ML is a subset of AI and the science of training machines to analyze and learn from data. Using algorithms, ML can detect patterns within data and create mathematical models to make predictions. The more access to data and experience, the more accurate the results. It’s often used in instances where the data or nature of the request/task changes frequently.

Large language models (LLMs)

LLMs are foundation models defined by Stanford Institute for Human-Centered AI as models trained on broad data. While LLMs can support a wide range of downstream tasks, generative AI is the more common use of output from foundation models. Newer models, known as multimodal models, combine more than one capability and understand several types of content. However, even single-mode use cases of LLMs, such as completing an unfinished song, are growing more flexible, adaptable, and powerful.

Small language models (SLMs)

SLMs provide most of the same capabilities found in LLMs but in a smaller size with training in smaller amounts of data. They are designed to handle simpler tasks and are generally more accessible and easier to use for organizations with limited resources. You can fine-tune them to meet your specific needs and run them on less expensive hardware such as your own PC. This is in contrast to LLMs, which require a large datacenter.

Mechanics of generative AI models

Generative AI models simulate how humans think through algorithms that "learn" to generate better output during a training phase. Generative AI then understands the information and creates content independently, usually with LLMs built with neural networks sitting at the core. By connecting with these models' APIs, companies can build applications with easy interfaces for specific tasks, increasing their accessibility.

Generative AI applications

Generative AI offers several exciting use cases and a wide range of personal and commercial applications.

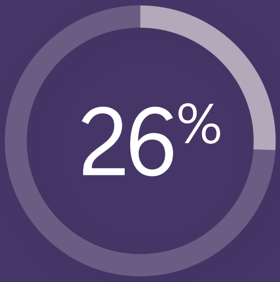
Type	Use cases	How it works
Text	Write news articles, poetry, and scripts, or translate text from one language to another.	The models can generate basic short-/medium-form writing, but are typically used for iteration or first drafts.
Images and videos	Produce film storyboards, product photos, magazine cover designs, and more.	New images can be generated based on existing ones, as well as upscaling and updating old content for higher definition media and new contexts.
Audio	Generate new music tracks as well as sound effects and voice acting.	Large datasets of audio recordings are collected and preprocessed. Using advanced machine learning models, new audio content is created in the form of music, speech, or sound effects.

Technology, algorithms, and models

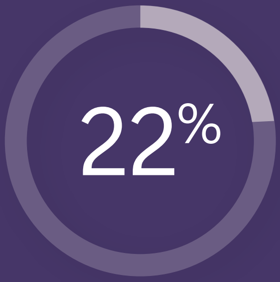
In this section, you'll learn foundational concepts about the technology and algorithms behind generative AI. With this understanding, you can assess the quality of AI-generated content, recognize security risks, and provide valuable feedback for improvement. This knowledge will empower you to advocate for responsible AI development, protect yourself online, and make better choices in an increasingly AI-driven world. In this section, learn about:

- Narrow and broad generative AI services
- Deterministic and nondeterministic security
- Prompt engineering
- Generative adversarial network (GAN)
- National Institute of Standards and Technology (NIST) Cybersecurity Framework
- AI shared responsibility model

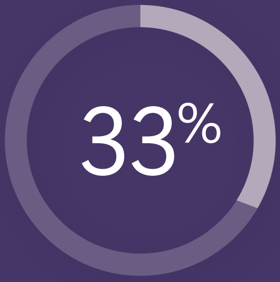
Enterprise applications of AI by the numbers



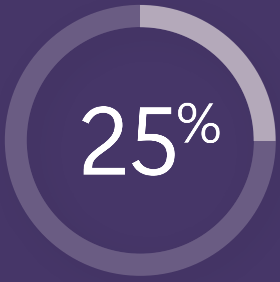
security and threat detection



automation of business processes



automation of IT processes



AI monitoring or governance

Source: IBM Global AI Adoption Index 2023

Narrow and broad generative AI services

Generative AI systems can be usefully divided into two groups: **narrow generative AI services** that are intended to assist a user with a finite and narrow range of tasks, and **broad generative AI services** that are open-ended by nature. The diagnostic test for narrowness is whether it would make sense to have a filter that rejects any requests (or responses) that are "off topic" with an error.

The differences between narrow and broad services

	Narrow services	Broad Services
What it is	Tailored to assist users with specific needs and limited tasks within well-defined boundaries	Open-ended approach capable of handling a wide range of tasks and inquiries
Area of expertise	Efficiency and precision, filtering out requests that fall outside the predefined scope	Addressing general questions and assisting with creating open-ended content, or information that allows for a wide range of responses
Examples	Customer service chatbots, spreadsheet automation tools, and software management aids	General inquiry chatbots like Bing Chat
Security considerations	Particularly valuable in security-sensitive environments where predictability and control are paramount	Robust security measures necessary to mitigate potential vulnerabilities and accommodate unexpected user behaviors

Deterministic and nondeterministic security risks

Deterministic security risks involve situations where vulnerabilities need to be prevented, detected, and patched. For example, consider if the model runs in a secure virtual machine, if the system's access to date goes through secure paths (authenticated and authorized), and what credentials it used. With deterministic security issues, AI is just software, and you should apply your existing understanding of software security.

Nondeterministic security risks are introduced via the AI system itself. For example, the possibility that the AI system might be deceived, confused, or coerced or that the user might misinterpret its results. These nondeterministic risks are new from the perspective of software security experts, as they cannot be "patched" in the way that traditional vulnerabilities are, they can only be made more or less likely to be triggered.

The differences between humans and generative AI in reviewing or generating content

Humans and generative AI bring different traits and aspects in reviewing or generating content. People excel in creativity, emotional intelligence, and contextual understanding. Generative AI can thrive in processing vast amounts of data and generating content at scale.

	Humans	Generative AI
Challenge	Humans are not immune to cyberthreats, such as social engineering. People can sometimes be misled.	AI can generate errors and be subject to malicious intent. It requires oversight and ongoing monitoring.
Training	Security awareness training can help people build trust over time. This can increase employees' scope of responsibilities accordingly.	Training AI, with techniques like metaprompts, metacognition, and fine tuning, can be adjusted and improved over time.
Reviews	A compliance function can encourage multiple crosschecks to minimize human error.	People in the decision-making loop can prevent generative AI errors from slipping through. You can also get AI systems to crosscheck each other.

Prompt engineering

Prompt engineering refers to the process of designing and refining prompts or instructions for AI models. It involves crafting input text that effectively communicates the desired task or query to the model. By carefully constructing prompts, users and developers can guide the model's behavior and improve its performance in various tasks, such as language generation, translation, or code completion.

Generative adversarial network (GAN)

The GAN is a leading type of generative model. This ML model offers a prominent framework for approaching generative AI. Using deep learning, GAN can create artificial data that resembles real data.

National Institute of Standards and Technology (NIST) Cybersecurity Framework

NIST Cybersecurity Framework is an AI risk management framework aimed at helping organizations identify generative AI risks and apply risk management techniques that match their priorities and objectives. This NIST Cybersecurity Framework is the result of input from more than 2,500 members. It focuses on 12 specific risks and provides developers with more than 400 actions to manage those risks.

AI shared responsibility model

AI responsibilities vary based on deployment model—software as a service (SaaS), platform as a service (PaaS), or infrastructure as a service (IaaS). The AI shared responsibility model details security considerations at each AI layer and highlights the importance of configuring before customizing. Responsibilities can shift based on deployment type, and below you can see an example using Microsoft as the provider.

This diagram illustrates responsibilities by deployment type.

		IaaS (BYO model)	PaaS (Azure AI)	SaaS (Copilot)
AI usage	User training and accountability	□	□	□
	Usage policy, admin controls	□	□	□
	Identity, device, and access management	□	□	◪
	Data governance	□	□	◪
AI application	AI plugins and data connections	□	□	◪
	Application design and implementation	□	□	■
	Application infrastructure	□	□	■
	Application safety systems	□	◪	■
AI platform	Model safety and security systems	□	◪	■
	Model accountability	□	◩	■
	Model tuning	□	◩	■
	Model design and implementation	□	◩	■
	Model training data governance	□	◩	■
	AI compute infrastructure	◪	■	■

■	Microsoft	◩	Model dependent
◪	Shared	□	Customer

Ethical and social implications

In this section, learn ways to address the ethical implications of generative AI, like its potential for misuse in deepfakes and misinformation, to foster trust with customers. Implementing ethical guidelines and regulations ensures responsible development and deployment, mitigating risks, and preserving societal values. In this section, learn about:

- Bias and fairness
- Misinformation
- Privacy
- Intellectual property
- Responsible AI
- Prompt filtering
- Metaprompt framework
- Retrieval augmented generation

Bias and fairness

Generative AI models can inadvertently inherit biases present in their training data. This can lead to unfair or discriminatory outcomes, particularly in areas like natural language processing where biased language is prevalent. Organizations worldwide are working on efforts to mitigate these biases.

Misinformation

When used incorrectly or unfairly, generative AI can be exploited to create false information, deep fakes, and fake news. Providing credibility and verifying sources is key to avoiding the spread of misinformation.

Privacy

The generation of realistic content can infringe on privacy rights. Deepfake videos, for example, can be used to impersonate individuals, potentially leading to identity theft and defamation.

Intellectual property

Determining ownership and copyright of content generated by AI systems can be a complex legal consideration as it’s determined who owns the output—the AI developer or the user.

Responsible AI

Responsible AI refers to the development and deployment of AI systems in a manner that prioritizes ethical considerations, fairness, transparency, accountability, and the wellbeing of individuals and society. It encompasses various principles and practices aimed at ensuring that AI technologies are developed and used in ways that align with societal values, legal frameworks, and human rights.

Prompt filtering

Prompt filtering is a method of detecting and rejecting inputs containing harmful or malicious intent and could circumvent the guardrails causing a jailbreak attack, or techniques used to bypass the safety measures and guardrails of an AI system.

Metaprompt framework

The metaprompt framework is designed to guide an AI system's behavior and improve system performance. Guardrails can be implemented to reduce the potential risk of unwanted behavior, such as the production of harmful content.

Retrieval augmented generation

Retrieval augmented generation is an architecture that augments the capabilities of an LLM like ChatGPT by adding an information retrieval system that provides grounding data. This can provide control over grounding data used by an LLM when it formulates a response, improving accuracy and reducing the risk of AI hallucinations.

Security and privacy for AI

Keep security and privacy top of mind when using generative AI, particularly concerning sensitive data and personal information, trust among users, and compliance with privacy regulations. Security for AI is defined as the security controls to help discover, govern, and protect against risks generated by generative AI development, deployment, and consumption. In this section, you'll learn about:

- Data privacy
- Data security
- AI compliance
- AI governance
- Zero Trust
- Security for AI
- Secure by design

Data privacy

Data privacy refers to the set of practices and concerns centered on the ethical collection, storage, and usage of personal information by AI systems.

Data security

Data security is a comprehensive approach to strengthening the security of your data. For best results, look for solutions that combine data and user context across your entire data estate, your devices, and your generative AI applications. For example, Microsoft data security solutions help you discover hidden risks to your data, protect and prevent data loss, and quickly investigate and respond to incidents.

AI compliance

AI compliance refers to the process of ensuring that AI-powered systems adhere to all applicable laws, regulations, guidelines, and best practices. Key aspects of AI compliance include:

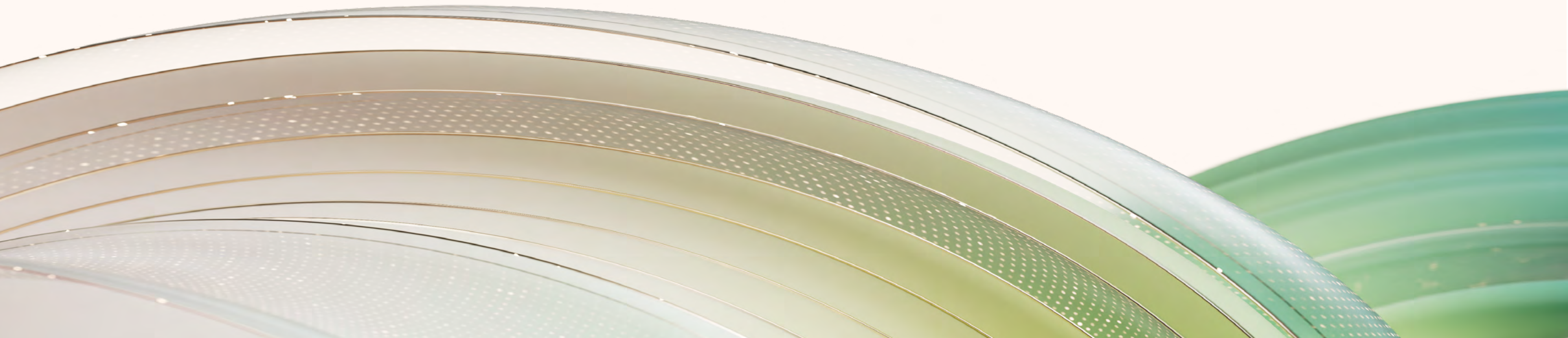
- Ensuring that AI systems comply with laws and regulations.
- Making certain that the data used to train AI systems is collected and used legally and ethically.
- Enacting internal policies and best practices related to AI.

AI governance

AI governance is a responsibility framework that helps you use AI in a thoughtful and compliant way.

Zero Trust

Zero Trust is a security model that helps boost security through the principles of "verify explicitly," "use least-privilege access," and "assume breach." This means authenticating and authorizing on all available data points, including user identity and service or workload, which limits user access.



Security for AI

Security for AI is focused on protecting yourself from threats aimed at AI use. One way to do this is through red teaming, a process that simulates the behavior of cybercriminals to identify vulnerabilities, weaknesses, and potential attack vectors within a system or network. This helps organizations proactively assess and improve their cybersecurity posture before adversaries can exploit vulnerabilities. AI red teaming typically involves:

- Planning and scoping to define the team's objectives.
- Simulation of attacks like phishing, malware injections, and network intrusion.
- Assessment of the team's detection and response.
- Reporting and analysis to document exercise findings.

Secure by design

Secure by design, in software engineering, means securing software and its capabilities from the outset. Security strategies are a key part of the development process and are built into the software design. At Microsoft, we use a secure-by-design method known as the Security Development Lifecycle (SDL). This approach can be applied to everything from classic waterfall to modern DevOps approaches. SDL practices include, for example: secure the software supply chain, define and use cryptography standards, and perform security testing.

Challenges and risks

Recognize the challenges and limitations of generative AI in order to make clear, informed decisions and have realistic expectations about integrating generative AI into your organization. In this section, learn about issues like data bias, realistic content generation, and the risk of overfitting or undesirable outputs:

- Data leakage
- Cross-prompt injection attacks (XPIA) (a.k.a. indirect and poisoned content)
- User-prompt injection attacks (UPIA) (a.k.a. direct content)
- Prompt leakage
- Overreliance
- AI hallucination
- Policy misalignment
- Shadow AI

Data leakage

Data leakage happens when information external to the training dataset is used to develop a model. Organizations often worry that data accessed by the generative AI service might "leak out." Three main concerns are that sensitive data might be included in outputs marked nonsensitive, that users might gain access to data they aren't entitled to, or that information may leak across users in a multiple-user environment. For example, data from one session might be regurgitated in a later session.

Cross-prompt injection attack (XPIA)

XPIA occurs when an attacker embeds a malicious payload inside external data. This typically involves hidden instructions the user may not be aware of producing a prompt that will confuse the model into treating attacker-controlled commands as though they were issued by the user. Two terms are important to understand when discussing XPIA:

- "Indirect content" means the malicious payload was added via other content rather than via the prompt. It could be a document the user references when asking "please summarize this document" or it could be a website when asking "please provide a review of this webpage."
- "Poisoned content" refers to the idea that the user may believe a document is legitimate, but an attacker hid malicious instructions within the text or image. This content could include video or audio.

User-prompt injection attack (UPIA)

A UPIA, also called a jailbreak attack, happens when someone intentionally exploits the vulnerabilities of an LLM-powered system. This technique can cause guardrails (mitigations) to fail and is often associated with other attack techniques like prompt rejection, evasion, and model manipulation. The objective typically is to overcome the security and safety boundaries for an unauthorized objective. An example is inducing the AI to provide detailed instructions on illegal activities.

Prompt leakage

Prompt leakage refers to the unintentional exposure of sensitive or confidential information through the input provided to an AI system. In the context of language models, prompt leakage occurs when a user inadvertently includes confidential details or secrets in their query or prompt, which the model then incorporates into its response that might be seen by other user or in content that might be shared with others. Unfortunately, prompt leakage is not possible to prevent, so the key is to never put anything into a metaprompt that you wouldn't want exposed to others outside your organization.

Overreliance

Overreliance is a type of system failure rather than an external attack. This can occur when the user trusts the output and believes it to be factual when in fact it is not accurate. There are several types of overreliance:

- Naïve: A user does not realize they shouldn't take what the AI says as factual.
- Rushed: A user is in a hurry and doesn't check the answer carefully.
- Forced: A user can't check the outputs due to design limitations.
- Motivated: A user blames the AI outcome as a justification for their actions.

The consequences of overreliance vary widely, from less serious when writing ad copy to more serious when diagnosing a medical issue. This can be managed with proper system design to ensure the user can remain aware and responsible for checking the outcome of the AI response.

AI hallucination

Originating in the AI research community, hallucination refers to the phenomenon of LLMs sometimes generating responses that are factually incorrect or incoherent. It can be useful as a means of generating new content like images, poems, and videos, but can be a problem if the user was expecting an accurate and factual answer. There are two categories of AI hallucination:

- False positives: The system makes up the answer without grounded information.
- False negatives: The system drops critical information from the grounded truth.

Policy misalignment

Some users of generative AI may have specific policies or goals that are not broadly applicable. For example, a financial services or pharmaceutical firm may have specific rules about what kinds of advice it can or cannot give, or a user may want to avoid certain language or terminology. Policies, especially when expressed in natural language (as in a metaprompt or a metacognitive loop), are prone to not capturing precisely what the authors intended, leading to policy misalignment, and often stark consequences.

Shadow AI

Shadow AI typically refers to AI systems or algorithms that have been incorporated without proper governance practices or security oversight. This may occur when employees use unsanctioned AI or integrate AI in ways that have not gone through rigorous testing and governance. This bypassing of established protocols or procedures can raise concerns about data privacy, security, and ethical implications.

Learn more about generative AI with Microsoft

We hope this guide has helped better prepare you for implementation of generative AI and realizing its full potential securely, compliantly, and confidently. Generative AI can have incredible transformative benefits in your organization, but it's ever-changing. For ideas on how to use generative AI in your organization, consider the most common use cases as documented by our e-book: AI Use Cases for Business Leaders.

→ **Learn more about Microsoft AI**