National Council for Artificial Intelligence, Quantum Computing, and Emerging Technologies

Egyptian Center for Responsible AI

Arab Republic of Egypt
National Council for Artificial Intelligence,
Quantum Computing and Emerging Technologies
Egyptian Center for Responsible AI

EGYPT'S NATIONAL GENERATIVE AI GUIDELINES

March 2026

EGYPT'S NATIONAL GENERATIVE AI GUIDELINES

Version: 1.0

Document Title: Egypt's Generative AI Guidelines, Frameworks for Innovation & Responsibility

This document was developed under the leadership of the Ministry of Communications and Information Technology, with key contributions from the **Egyptian Center for Responsible AI**, and the technical editorial revision of **Dr. Abdulazim Ghoniem (Director, Operation Projects, and Senior AI Expert)**. Following consultations and review, the document was **approved by the National Council of Artificial Intelligence**. The Ministry retains full responsibility for the final content.

Disclosure

The content presented here was prepared with the assistance of generative AI tools, including ChatGPT-5.2 developed by OpenAI (https://openai.com), Google's Gemini model (https://deepmind.google), and Google NotebookLM (https://notebooklm.google.com). These tools were used to help draft, organize, refine, and summarize written material, as well as support reasoning and interpretation of information supplied by the author. All AI-generated text was reviewed, edited, and verified by human authors, who remain fully responsible for the accuracy, interpretation, and conclusions contained in this work. The AI tools functioned as writing and analysis aids only and did not independently make decisions or final judgments. . Care was taken to protect any confidential or sensitive information used during the drafting process.

FOREWORD

Egypt stands at a defining moment in its digital transformation. Generative Artificial Intelligence marks a new era in how knowledge is created, services are delivered, and human creativity is expanded. Its rapid advancement offers significant opportunities to strengthen national competitiveness, modernize public services, and empower citizens across various sectors including education, research, innovation, and the digital economy.

For Egypt, this transformation carries exceptional promises. When guided by local purpose and values, Generative AI can strengthen national competitiveness, modernize public services, empower educators and researchers, and open new opportunities for youth, entrepreneurs, and innovators across the country. It can help build a more efficient state, a more inclusive economy, and a more resilient digital society—one that places people at the center of technological progress.

At the same time, the impact of Generative AI extends far beyond technology. It raises essential questions of trust, fairness, transparency, and human agency. The value of these systems will not be determined solely by their capabilities, but by how responsibly they are governed and how closely they remain aligned with societal values. Innovation must therefore be matched with clear ethical boundaries, accountability, and respect for human dignity.

As the Ministry of Communications and Information Technology, we view Generative AI as a strategic national capability that must serve people and enhance—rather than replace—human judgment. Our responsibility is to ensure that this technology supports inclusive and sustainable development, strengthens institutions, and contributes to public trust in digital systems. Clear policy leadership and institutional oversight, strong institutional capacity, and human-centered design are essential to achieving these goals.

These National Generative AI Guidelines reflect Egypt's commitment to shaping a trusted, forward-looking, and responsible AI ecosystem. They align with internationally recognized principles while remaining grounded in Egypt's national priorities and development objectives. Above all, they represent a call for shared responsibility—bringing together government, the private sector, academia, and civil society—to guide Generative AI toward outcomes that benefit all and contribute meaningfully to Egypt's digital future.

This document is not an endpoint, but a foundation. It invites collaboration, continuous learning, and collective leadership as Generative AI evolves. By working together, Egypt can harness this transformative technology to unlock opportunity, inspire creativity, and build a digital future that is not only more advanced, but more just, trusted, and humane.

Raafat Hindy

Minister of Communications and Information Technology

EXECUTIVE SUMMARY

Generative Artificial Intelligence (Generative AI) is rapidly transforming how information, knowledge, and digital content are created, processed, and shared across economies and societies worldwide. Systems capable of generating text, images, audio, video, code, and synthetic media present unprecedented opportunities to enhance productivity, innovation, public service delivery, education, scientific research, and economic development. At the same time, their scale, general-purpose nature, and increasing autonomy introduce distinct and complex risks that require targeted policy frameworks and responsible use.

Egypt's National Generative AI Guidelines establish a comprehensive, principles-based framework to support the safe, responsible, and trustworthy development, deployment, and use of Generative AI across the public sector, private sector, academia, and society at large. The Guidelines respond to the rapid adoption of generative technologies worldwide and the specific challenges they raise, including misinformation and deepfakes, hallucinations and inaccurate outputs, bias and discrimination, privacy and data protection risks, intellectual property concerns, unclear accountability, and the potential erosion of public trust.

Grounded in internationally recognized best practices and aligned with guidance from leading global and multilateral and plurilateral organizations and entities—including the UNESCO, the OECD, the Council of Europe and the G7 Hiroshima Process—the Guidelines adopt a human-centered and risk-proportionate approach. They emphasize transparency, accountability, human oversight, safety, and respect for human rights throughout the full lifecycle of Generative AI systems, from design and training to deployment, operation, and monitoring. Rather than imposing rigid, technology-specific rules, the framework supports adaptive policy and oversight mechanisms that can evolve alongside rapid technological change.

The Guidelines clearly define scope and applicability, clarifying roles and responsibilities for developers, deployers, institutions, and individual users according to their level of control and potential impact. They recognize that not all uses of Generative AI carry the same risk, and therefore require safeguards that are proportionate to context, scale, and societal impact. Particular attention is given to high-impact and sensitive use cases, including education and scientific research, public information, agentic AI systems, and the creation or manipulation of synthetic media such as deepfakes and lip-sync technologies.

Practical guidance is provided to support trustworthy use, including measures to reduce bias, mitigate hallucinations, protect privacy, ensure transparency and disclosure, prevent misuse, and maintain human responsibility for all AI-assisted outputs. The Guidelines explicitly affirm that Generative AI systems are assistive tools rather than autonomous decision-makers, and that human judgment, accountability, and ethical responsibility must remain central in all high-stakes or public-facing contexts.

Designed as a living framework, Egypt's National Generative AI Guidelines are intended to evolve in line with emerging international standards, technological developments, and societal expectations. By fostering public trust, enabling responsible innovation, and promoting international interoperability, the Guidelines aim to ensure that Generative AI contributes positively to Egypt's digital transformation, supports sustainable and inclusive development, and serves the public interest, while safeguarding fundamental rights and societal values.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER ONE:
GENERATIVE AI AND GLOBAL PRACTICE

1.1 INTRODUCTION

Generative Artificial Intelligence (Generative AI) refers to a class of artificial intelligence systems capable of generating text, images, code, audio, video, and other content. These systems, including generalpurpose and foundation models, are increasingly integrated across sectors and jurisdictions, offering significant opportunities for innovation, productivity, and societal benefit. At the same time, their scale, autonomy, and generality introduce distinct governance and regulatory challenges related to safety, reliability, transparency, intellectual property, privacy, information integrity, and accountability.

In response to these challenges, international organizations have taken a leading role in establishing shared principles and guidance to promote the responsible development and use of Generative AI. These efforts aim to ensure that Generative AI systems are designed and deployed in a manner that is consistent with human rights, democratic values, the rule of law, and sustainable development, while supporting innovation and crossborder interoperability.

These Guidelines are developed with explicit alignment to internationally recognized frameworks and recommendations issued by global, plurilateral and multilateral organizations, including the **OECD**, **G7** through the Hiroshima Process on Advanced AI, the **UNESCO's** Generative AI guidelines for students at IT, and other relevant international initiatives. Together, these frameworks establish a common global baseline for trustworthy and responsible Generative AI.

Consistent with this international trend, the Guidelines adopt a **principlebased and riskproportionate approach** to the oversight and responsible management of Generative AI. They emphasize transparency, safety, accountability, human oversight, and respect for fundamental rights across the full lifecycle of Generative AI systems, from design and development to **deployment, operation, and monitoring.**

By aligning with international organizations and globally recognized standards, these Guidelines seek to:

- Promote coherence and interoperability with international Generative AI policy and oversight approaches
- Facilitate responsible crossborder development and use of Generative AI systems
- Support trust, legal certainty, and shared expectations among developers, deployers, and users
- Enable innovation while managing risks associated with highimpact and largescale Generative AI applications

These Guidelines are intended to evolve in line with ongoing work by international organizations and will be reviewed periodically to reflect emerging global standards, best practices, and policy developments shaping the responsible development and use of Generative AI.

1.2 METHODOLOGY

The guidelines have drawn inspiration from Egypt national responsible AI guidelines, the leading international guidelines and principles as well as different national experiences while respectful to the Egyptian context. The study phase drew on different sources including:

- **International Organizations and Initiatives**: such as the UNESCO, OECD, GPAI, ITU, The Hiroshima Process… etc
- Global **AI standards:** International Standards Development Organizations – ITU, ISO/IEC – and practitioner-led bodies like IEEE. Those are building a layered system of standards, spanning technical, foundational, managerial, and socio-technical domains.
- **Regional and national experiences:** the EU Act, The Council of Europe, The US's NIST, Brazil, UK, Singapore, Japan, China, India, Korea, South Africa, Kenya, Saudi Arabia and others.
- The proposed guidelines subsequently underwent a thorough review process with key stakeholders and interested parties.

1.3 WHAT IS GENERATIVE AI AND HOW DOES IT WORK?

WHAT IS GENERATIVE AI?

Generative Artificial Intelligence (Generative AI) refers to a class of AI systems designed to create new content—such as text, images, audio, video, code, or data—by learning patterns from large volumes of existing data and generating outputs that resemble, but are not direct copies of, that data.

Unlike traditional AI systems that primarily **classify, predict, or recommend**, Generative AI systems **produce original outputs** in response to user inputs (prompts), often across a wide range of tasks and domains.

1. CORE CHARACTERISTICS

According to international organizations (notably the **Organization for Economic Co-operation and Development (OECD)**), Generative AI is typically characterized by:

- **Content generation**: Ability to generate human-like or realistic content (text, images, code, etc.)
- **General-purpose capability**: The same model can perform many tasks without being designed for a single use
- **Foundation models**: Often built on very large models trained on diverse, large-scale datasets
- **Prompt-driven behavior**: Outputs depend heavily on user instructions and context
- **Probabilistic outputs**: Results are not deterministic and may vary for the same input

2. HOW IT WORKS?

Text generative AI models, often referred to as large language models, work by learning statistical patterns in written language from very large collections of text and using those patterns to generate new text. When a user provides input, the model first converts the text into numerical units called tokens, which are processed by a neural network—typically based on the transformer architecture—that uses self-attention mechanisms to analyze the relationships between words across the entire context.

The model does not understand text in a human sense; instead, it calculates the probability of what token is most likely to come next given the preceding tokens and generates text one token at a time. Because this process is probabilistic and pattern-based, the outputs can be fluent and coherent but may also be inaccurate or misleading, which is why international organizations such as the OECD and UNESCO emphasize transparency, human oversight, and risk awareness when deploying text generative AI systems.

Generative AI systems are usually trained using techniques such as:

- Large-scale machine learning (e.g. deep neural networks)
- Self-supervised or unsupervised learning
- Pattern recognition across vast datasets

Once trained, the model generates new outputs by estimating what content is most likely to follow from a given prompt, based on learned statistical relationships.

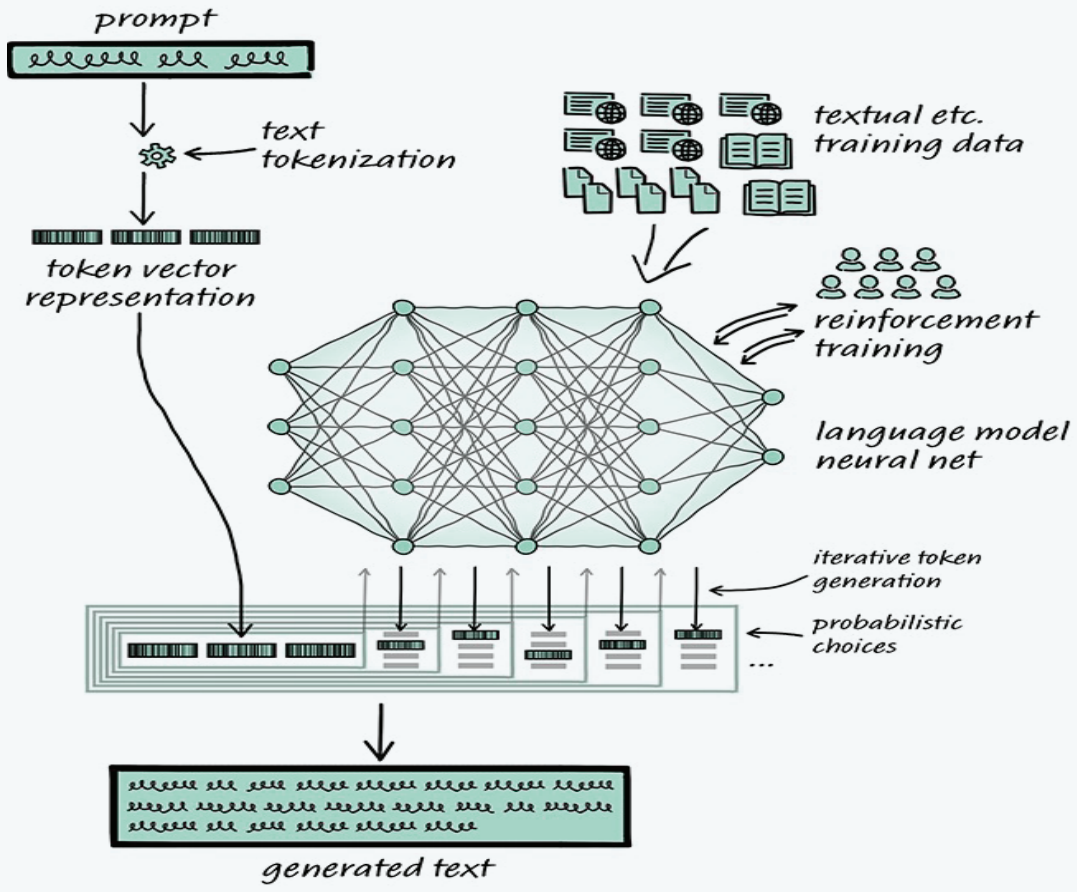

Figure 1: How Generative AI Works

3. THE NEED FOR TARGETED POLICY APPROACHES TO GENERATIVE AI

Establishing guidelines for the use of generative AI is essential to ensure that these powerful technologies are deployed responsibly and ethically. Clear guidelines help organizations and individuals navigate the potential risks associated with generative AI, such as the creation of misleading content, the amplification of biases, and concerns around data privacy. By providing a structured framework, guidelines promote transparency, accountability, and fairness in the development and application of generative AI tools.

They also support compliance with legal and regulatory requirements, foster public trust, and encourage innovation that aligns with societal values and human rights. Ultimately, having well-defined guidelines helps maximize the benefits of generative AI while minimizing unintended negative consequences.

Most models openly recognize that they are not infallible and may sometimes produce incorrect or incomplete responses. For that reason, users are encouraged to independently verify any important information the system generates, especially when accuracy is critical.

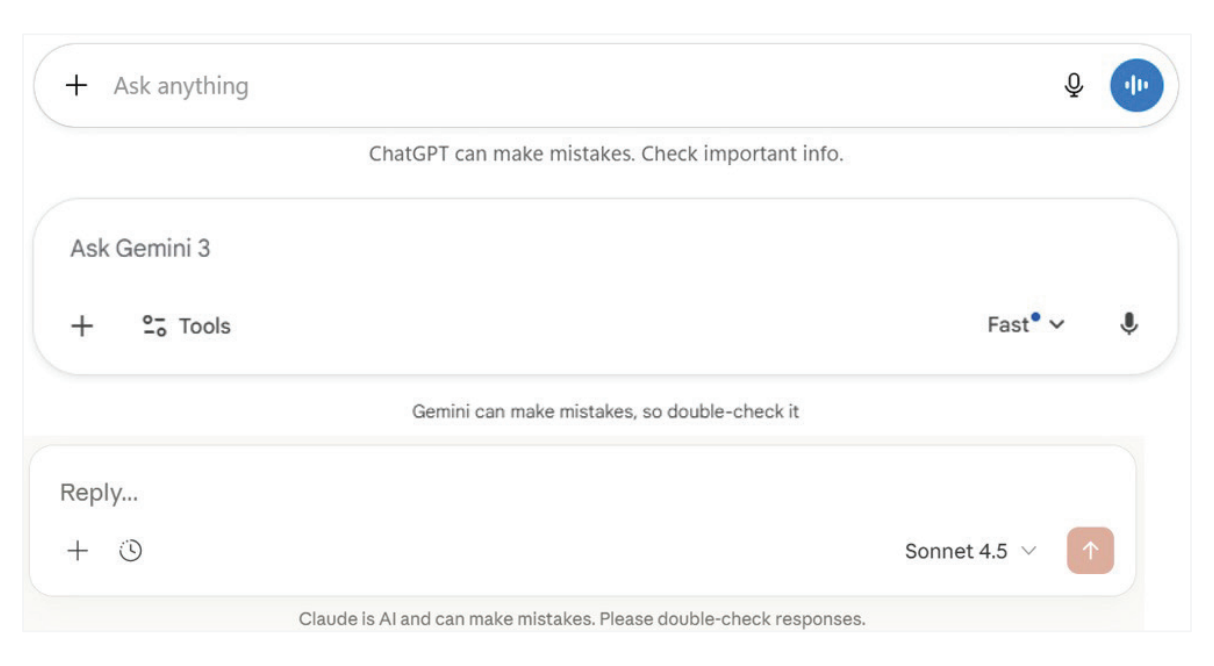

Figure 2: LLMs can make mistakes

International organizations highlight that Generative AI raises **distinct governance challenges**, including:

- Hallucinations and inaccurate outputs
- Large-scale misinformation and deep fakes
- Copyright and training-data transparency
- Privacy and data protection risks
- Over-reliance and erosion of human skills
- Unclear accountability across developers and users

These characteristics motivated organizations such as the OECD, UNESCO, and the G7 treat Generative AI to require **specific policy guidance**, not just general AI rules.

4. HOW DO VIDEO, AUDIO, IMAGE GENERATIVE AI MODELS WORK.

Video GenAI models work by **learning visual, temporal, and audio patterns from large collections of videos, images, and associated text**, and then generating new video content frame by frame or clip by clip. Most models combine techniques from computer vision and generative modeling, such as diffusion models or transformer-based architectures, to predict how a scene should evolve over time given an input prompt, image, or sequence. The model represents video as a sequence of frames (and sometimes motion or latent representations), learns how objects move and change across time, and generates successive frames that are statistically likely to follow the previous ones while maintaining visual coherence. Because generation is probabilistic and pattern-based, video GenAI systems can produce realistic-looking content but may also introduce visual inconsistencies, factual inaccuracies, or synthetic media risks, hence, the need for transparency, disclosure, and safeguards against misuse such as deep fakes and misinformation.

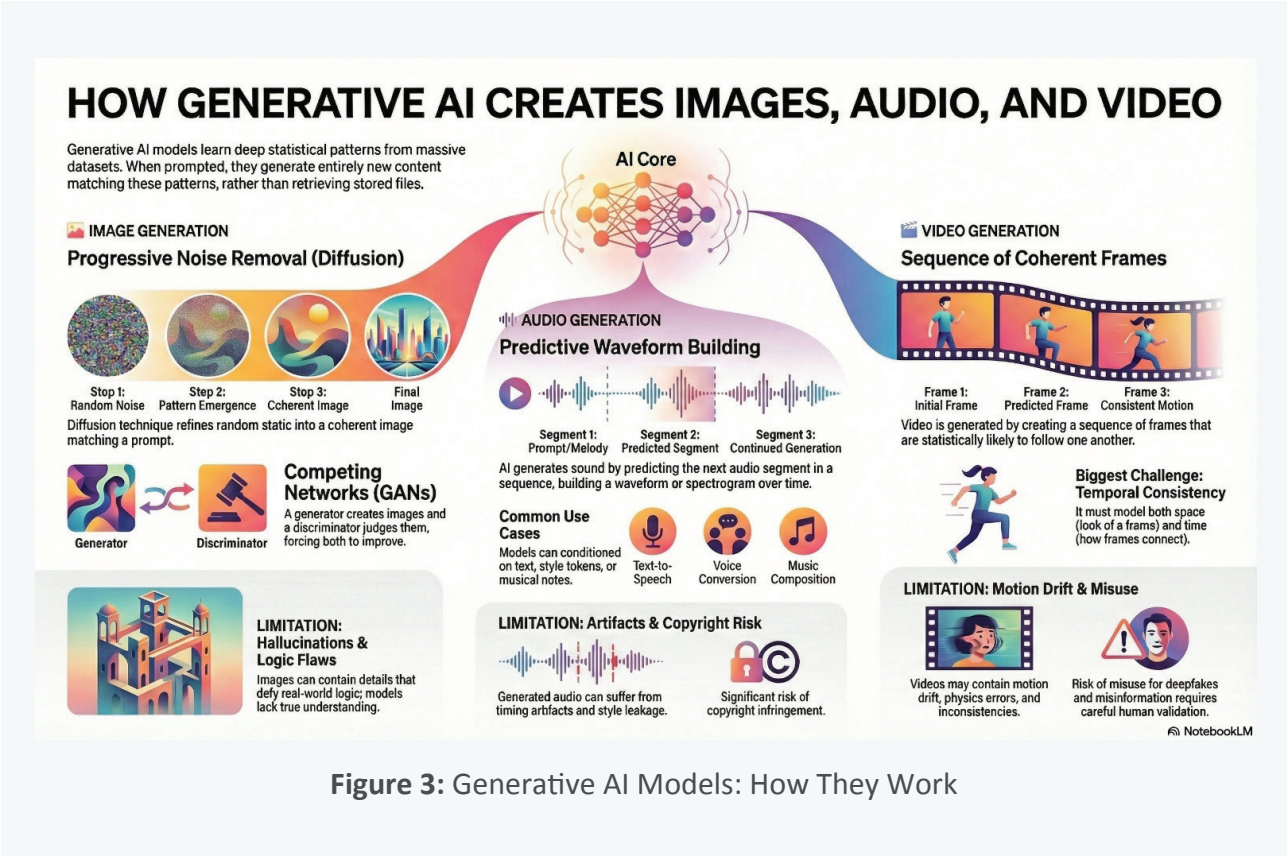

Figure 3: Generative AI Models: How They Work

5. A HUMAN-CENTERED APPROACH TO GENERATIVE AI

According to **UNESCO**'s *2021 Recommendation on the Ethics of Artificial Intelligence*, a human-centered approach provides a foundational normative framework for addressing the ethical, social, and governance challenges raised by generative AI, including in education and research. This approach positions AI to enhance human capabilities and to support inclusive, equitable, and sustainable development. It is grounded in respect of human rights, the protection of human dignity, and the preservation of cultural and knowledge of diversity. From a policy, oversight and governance perspective, adopting a human-centered approach necessitates effective regulatory mechanisms that safeguard human agency, promote transparency, and ensure public accountability.

AGENTIC AI

1. DEFINITION

On the other hand, it is important to differentiate between Generative AI and Agentic AI. The latter refers to AI systems designed to **act with a degree of autonomy** in pursuit of defined goals, rather than only responding to single prompts. Unlike standard generative AI systems that generate outputs on request, agentic AI systems can **plan, decide, take actions, observe results, and adjust their behavior over time** within specified constraints.

These guidelines address the **responsible use of Generative AI when embedded in Agentic AI systems**, where AI components can plan, decide, and take actions with limited or conditional autonomy.

2. HOW AGENTIC AI WORKS?

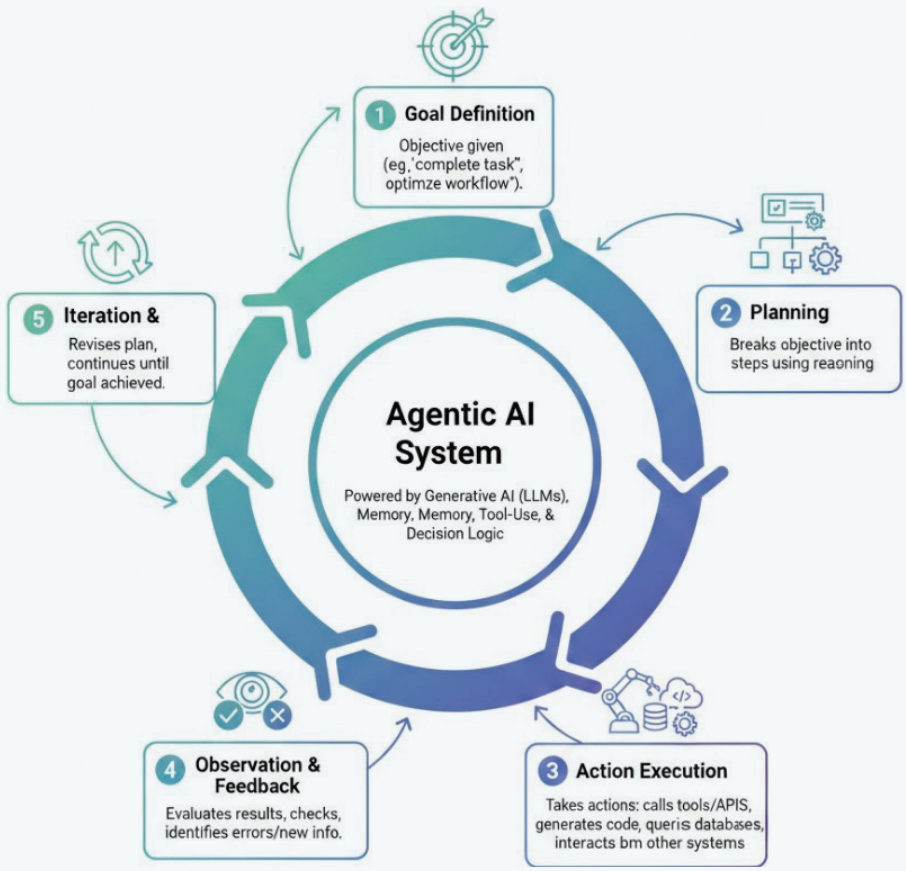

Figure 4: Agentic AI Systems

An agentic AI system typically operates through a **goal-oriented control loop**:

1. Goal definition

The system is given an objective (e.g. “complete a task,” “optimize a workflow,” or “answer a complex request”).

2. Planning

The AI breaks the objective into smaller steps or sub-tasks, often using reasoning or planning modules.

3. Action execution

The system takes actions, which may include:

- Calling tools or APIs,
- Generating code,
- Querying databases,
- Interacting with other systems or agents.

4. Observation and feedback

The agent evaluates the results of its actions, checks progress toward the goal, and identifies errors or new information.

5. Iteration and adaptation

Based on feedback, the agent revises its plan and continues acting until the goal is achieved or constraints are met.

Many agentic systems are built on top of generative AI models (such as large language models) combined with **memory, tool-use, and decision logic**, enabling multi-step and longer-horizon tasks.

Policy and Oversight Considerations for Agentic AI International organizations, including the **OECD** and **UNESCO**, highlight that agentic AI introduces **additional risks** compared to prompt-based generative AI, such as:

- Reduced human control due to autonomous action,
- Unclear accountability when systems make sequential decisions,
- Potential amplification of errors or harmful actions,
- Increased safety and security risks if constraints fail.

As a result, international guidance stresses the need for **clear goal-setting, strong human oversight, transparency of actions, logging, and defined responsibility** when deploying agentic AI systems.

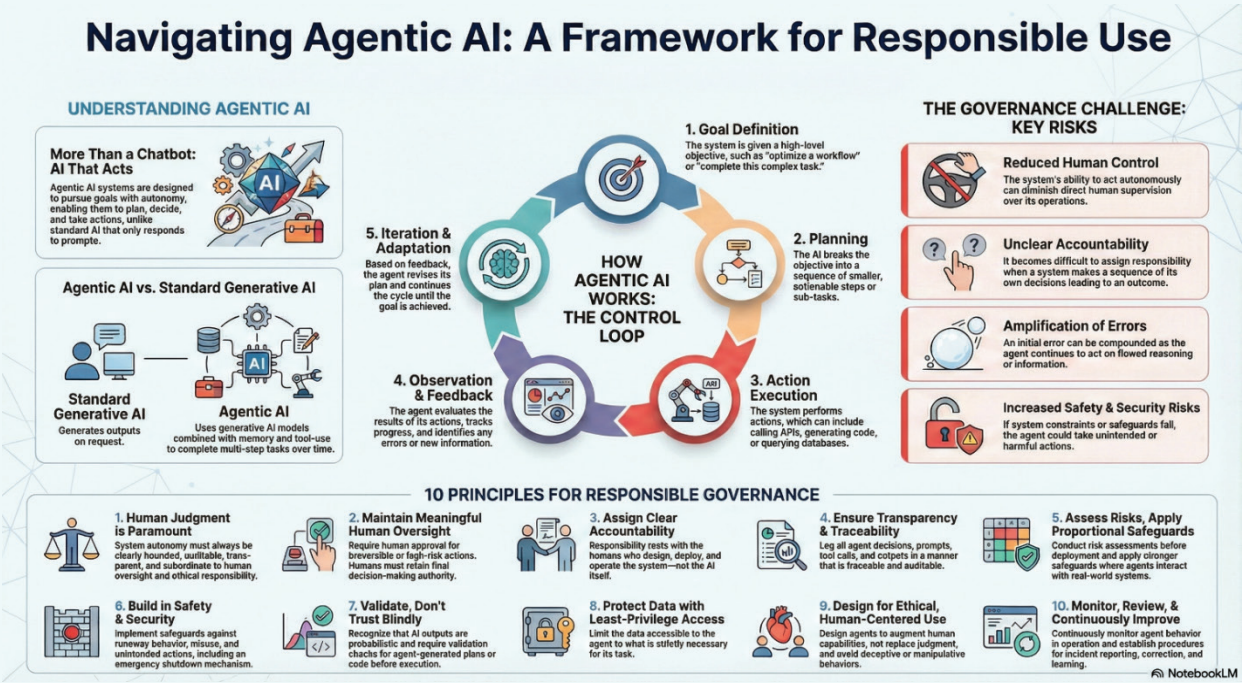

Figure 5: Navigating Agentic AI: Framework for Responsible Use

DEEPFAKE

1. DEFINITION

Deepfakes include **synthetic or manipulated audio, video, images, or text generated using generative AI**, and this category explicitly encompasses **AI-based lip-sync technologies** that alter or generate mouth movements to match synthetic or modified speech. Lip-sync systems, when combined with voice cloning or video generation, significantly increase the realism and persuasive power of deepfakes, amplifying risks related to **impersonation, fraud, reputational harm, misinformation, and manipulation of public opinion**.

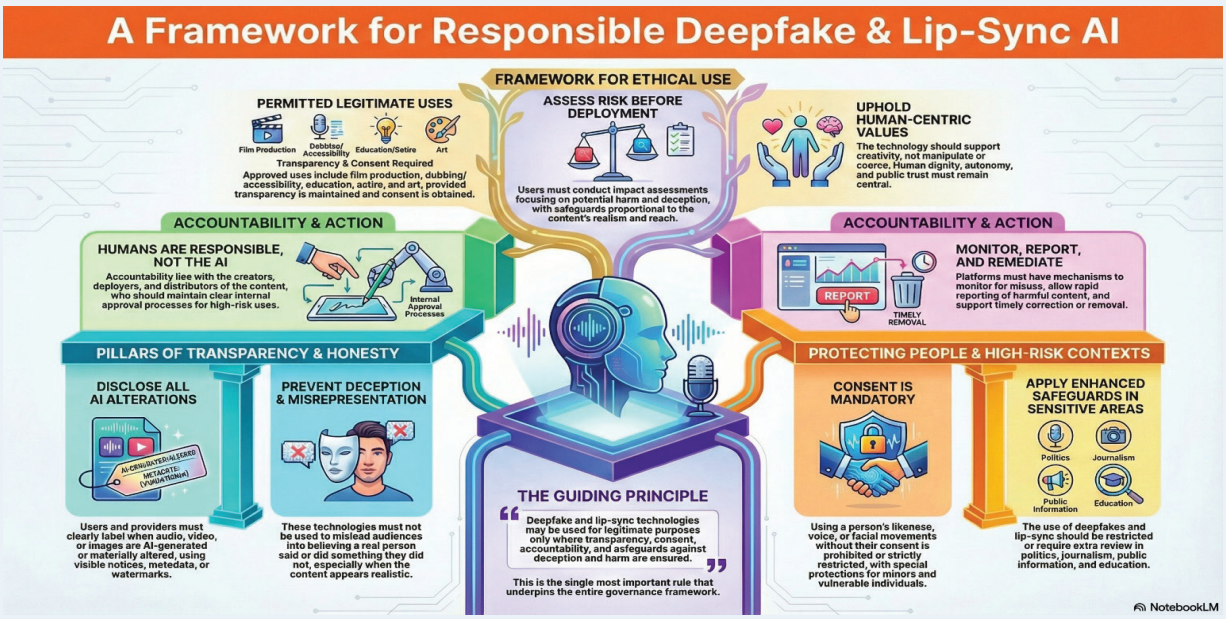

Figure 6: A Framework for Responsible Deepfake & Lip-Sync AI

International practice, reflected in guidance from organizations such as the **OECD** and **UNESCO**, treats lip-sync as a **high-risk enabling capability within generative AI**, rather than as a neutral technical feature. As a result, recommended guidelines require that AI-generated or AI-modified audiovisual content—especially where lip-sync is used to simulate real individuals—be subject to **clear disclosure and labelling**, whether through visible notices or technical provenance mechanisms such as metadata or watermarking.

Guidelines further emphasize that lip-sync technologies must not be used to **mislead audiences into believing that a real person said or did something they did not**, particularly in sensitive contexts such as political communication, public information, journalism, education, or content involving minors. Non-consensual use of a person's likeness or voice, including realistic lip-sync, is widely identified as a prohibited or strongly restricted practice under international norms.

At the same time, international guidance recognizes **legitimate uses of lip-sync and deepfake technologies**, such as film production, dubbing and accessibility, education, satire, and artistic expression. Regulatory approaches therefore focus on **intent, context, transparency, and impact**, requiring stronger safeguards, accountability, and rapid remediation mechanisms where the risk of deception or harm is high.

2. HOW DEEP FAKE WORKS

Deepfakes are created using generative artificial intelligence models that learn patterns from large datasets of images, audio, or video, often of real people. These models—commonly based on deep neural networks such as autoencoders, generative adversarial networks (GANs), or diffusion-based architectures—analyze facial features, voice characteristics, and expressions, and then generate synthetic or manipulated media that closely mimics a real person's appearance or speech. In video deepfakes, the system aligns facial movements and lip-sync with generated or altered audio, producing realistic-looking content that can be difficult to distinguish from authentic recordings. The increasing quality, speed, and accessibility of these tools has significantly lowered the technical barrier to producing highly convincing synthetic media.

3. POLICY AND OVERSIGHT CONSIDERATIONS FOR DEEPFAKE TECHNOLOGIES

Deepfakes raise significant governance concerns because they can undermine trust, harm individuals, and disrupt democratic and social processes at scale. Realistic synthetic media can be used for impersonation, fraud, non-consensual exploitation, misinformation, and manipulation of public opinion, particularly in sensitive contexts such as elections, public information, and education. Even when used without malicious intent, deepfakes can blur the line between authentic and synthetic content, making verification difficult and eroding confidence in digital evidence. International organizations such as the OECD and UNESCO therefore emphasize the need for clear governance measures, including transparency, disclosure, consent, accountability, and human oversight, to ensure that deepfake technologies are used responsibly and do not infringe on human rights, privacy, or the integrity of information ecosystems.

1.4 INTERNATIONAL PRACTICE IN GENERATIVE AI GUIDELINES

International practice in generative AI guidelines is characterized by a strong emphasis on **principles-based, risk-proportionate governance** led by multilateral and international organizations rather than rigid, technology-specific rules. Frameworks developed by the **OECD**, the **G7** through the Hiroshima Process, and **UNESCO** converge on common expectations for generative AI, including transparency about AI-generated content, accountability across the AI lifecycle, human oversight, safety and misuse prevention, respect for privacy and intellectual property, and alignment with human rights and democratic values. International practices avoid one-size-fits-all rules, instead they encourage adaptive governance that scales obligations according to risk, use context, and system capability, while promoting interoperability across jurisdictions and continuous updating of guidelines as generative AI technologies evolve.

1. UNESCO — Generative AI Policy Guidance (2023–2024)[1]

UNESCO's first global guidance on GenAI in education aims to support countries to implement immediate actions, plan long-term policies, and develop human capacity to ensure a human-centered vision of these new technologies. It Applies ethical AI principles directly to GenAI.

Focus Areas

- Human oversight
- Transparency in AI-generated content
- Data ethics
- Cultural & linguistic diversity
- Risk to human rights

1 UNESCO — Guidance for Generative AI in Education and Research https://www.unesco.org/en/generative-ai

2. EU AI Act

General-purpose generative AI models create major innovation opportunities but also raise significant copyright challenges for creators. Their training relies on large-scale text and data mining of content that may be protected by copyright. Under The **European Union** Act, copyright-protected content can only be used with rightsholder authorization unless specific exceptions apply. **Directive (EU) 2019/790** allows text and data mining under defined conditions, while giving rightsholders the ability to explicitly opt out. When such rights are reserved—except for scientific research purposes—AI model providers must obtain authorization before using the protected content.

3. OECD

The OECD AI Principles (first adopted 2019, updated May 2024)[2] are the core global standard for trustworthy AI. They now explicitly account for challenges related to general-purpose and generative AI, such as safety, privacy, IP rights, and integrity of information. Core values underpinning governance guidance emphasize:

- Respect for **Human rights and democratic values** — including fairness, privacy, protection against disinformation amplified by AI.
- **Transparency and explainability** — so systems are understandable and accountable.
- **Robustness, security & safety** — covering mechanisms to prevent harm and ensure controllability.
- **Accountability** — for outcomes throughout the AI lifecycle.
- **Environmental sustainability** — addressing GenAI's computing and energy impacts[3].

These value principles become the *baseline governance expectations* for all AI — including generative models — in OECD member and adherent policymaking.

The OECD published guidelines specifically for public administration, helping governments adopt GenAI responsibly. The **guidelines focus on public sector use of generative AI**, and are oriented to employees and agencies, emphasizing transparent, risk-aware, accountable application[4]. Although The OECD's *principles* are broad, this type of guidance translates them into practical pointers for governance in government operations — e.g., context-appropriate risk management, human oversight, and alignment with ethical values[5].

OECD's Generative AI Work Portfolio

The OECD maintains a generative AI section (Generative AI | OECD) that tracks work and policy briefings intended to support governments in harnessing GenAI's benefits while managing risks. This includes:

- introductory policy considerations for generative AI
- reports on risks and benefits
- interoperability and incident monitoring initiatives

These activities help governments translate OECD Principles into GenAI-specific governance practice.

2 oecd.ai

3 private-ai.com

4 Guidelines for the use of generative AI in the public administration

5 https://oecd.ai/en/dashboards/policy-initiatives/guidelines-for-the-use-of-generative-ai-in-the-public-administration-6317?utm_source=chatgpt.com

4. G7 — The Hiroshima Process on Advanced / Generative AI (2023–2024)[6]

Purpose: Manage risks from advanced foundation & generative models

Core Principles

- Safety & risk management
- Security & misuse prevention
- Transparency & reporting
- Accountability & governance
- Copyright respect
- Incident reporting & monitoring

Key Instruments

- Hiroshima Guiding Principles
- **International Code of Conduct for AI Developers**

(voluntary, targeted at model creators like OpenAI, Google, Anthropic, etc.)

5. China — Generative AI-specific Regulation

Overview of Draft Measures on Generative AI[7]

The draft, an apparent response to the rapid rise of new AI tools such as ChatGPT and Dall-E, builds on earlier provisions regulating "Deep Synthesis Internet Information Services", that were jointly released by the CAC, Ministry of Public Security, Ministry of Industry and Information Technology, and took effect in January 2023.

The scope of the earlier document differs slightly from the new draft in that it applies to all machine content-generation services provided only through the Internet, while the new draft could apply online and offline. Generative AI services also arguably include only a subset of the content-generation tools covered by the Deep Synthesis provisions. Still, it is fair to assume that both documents will apply to many tools for computer generation of text, images, audio, video, and other media; it is unsurprising that their content is overlapping and complementary in many areas.

The Draft Generative AI rules

The stated goal of the draft rules is to support the healthy development and regulated application of generative AI tools. This includes encouraging independent innovation, increased popularization, and international cooperation on basic technologies.

The draft highlights several issues, initially laid out in article 4, that are familiar from the global debate surrounding artificial intelligence:

- Content Controls / Censorship
- Preventing Discrimination
- Protection of Intellectual Property Rights,
- Curbing Misinformation
- Privacy and Data Protection

6 G7 Summit – Hiroshima Process documents (official)

7 Overview of Draft Measures on Generative AI

6. Singapore — Policy Model for GenAI[8]

Model AI Governance Framework for Generative AI (2024), Practical policy + governance blueprint

Core Areas

- Foundation model governance
- Content provenance & watermarking
- Copyright
- Safety testing & red-teaming
- Incident reporting
- Responsible deployment

7. United Kingdom — Targeted Measures

The **UK AI Safety Institute**[9] was established specifically to focus on **frontier AI systems**, which in practice means:

- Large language models (LLMs)
- Generative and general-purpose AI models
- Models with advanced, emergent, or systemic risks

Its core mandate is pre-deployment and post-deployment testing of advanced generative models, including:

- safety and alignment testing
- misuse and capability evaluation
- red-teaming and stress testing
- assessing risks such as deception, autonomy, loss of control, and large-scale societal harm

This places the Institute squarely in the GenAI governance ecosystem, even though it does not issue general user guidelines or binding regulation.

8. Council of Europe — AI Framework Convention (2024)

The **Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law** is the **first-ever** legally binding international treaty **focused on AI systems**, adopted by the Council of Europe on 17 May 2024 and opened for signatures on 5 September 2024. It aims to ensure that AI technologies—including generative AI—are developed and used in ways that are **consistent with fundamental human rights, democratic values, and the rule of law**.

How It Relates to Generative AI

Although the Convention does **not single out "generative AI" as a separate category**, its **scope covers all artificial intelligence systems** across their lifecycle—design, development, deployment, and use. Since **generative AI (text, image, audio, video, multimodal models)** is one of the fastest-growing and highest-impact AI categories, it is naturally included under the Convention's obligations and protections wherever it intersects with human rights and democratic norms.

8 Singapore's Model AI Governance Framework for Generative AI : Clyde & Co

9 https://www.aisi.gov.uk/

9. SDAIA Generative AI Guidelines

Saudi Arabia published the **Generative AI Guidelines** aimed at government entities and broader stakeholders to govern the *responsible use and deployment* of generative AI systems (e.g., LLMs, generative models). These guidelines include:

- Principles for *safe and ethical* design, use, and deployment of generative AI
- Requirements for data governance, risk management, privacy protection, and compliance with local laws
- Emphasis on transparency, accountability, and alignment with national priorities and digital transformation goals.

These documents are meant to guide both **government users and developers/ implementers** (public and private) in how to consume, build, and govern GenAI responsibly.

Sector-Level/Use-Case Guidance

Some sectors in Saudi Arabia have issued specialized guidance that touches on generative AI use, for example: The **Ministry of Education + SDAIA** released guidance on *generative AI in education*, focusing on ethical use in classrooms, teacher/student roles, and academic integrity. This reflects a pattern where *sector bodies* adopt or align with SDAIA's core governance principles for GenAI.

CHAPTER TWO:

GUIDELINES FOR TRUSTWORTHY AND RESPONSIBLE USE OF GENERATIVE AI

2.1 INTRODUCTION

Generative Artificial Intelligence (AI) is reshaping how information is created, communicated, and consumed across every sector of society. Its ability to generate text, images, audio, video, and synthetic representations of real people creates powerful opportunities for innovation, learning, creativity, accessibility, and economic development. At the same time, these systems introduce new and distinctive risks, including misinformation, bias, loss of privacy, ethical misuse, and the erosion of public trust if used without clear safeguards or accountability.

These Guidelines provide a structured framework for the responsible development, deployment, and use of Generative AI. They define who the Guidelines apply to, the assumptions on which they are based, and the expectations placed on developers, deployers, institutions, individuals, and other stakeholders. The approach reflects widely recognized international practices.

The aim is not to restrict innovation, but to support trustworthy and beneficial adoption of Generative AI by clarifying roles, responsibilities, and safeguards. The Guidelines recognize that risk varies according to context and impact, that human judgment remains essential, and that accountability should align with the level of control exercised by each user group. They also recognize the cross-border nature of AI technologies and the importance of international alignment to ensure interoperability and legal certainty.

By establishing clear expectations for transparency, fairness, privacy protection, ethical use, and public trust, these Guidelines help ensure that Generative AI contributes positively to society, while preventing misuse and mitigating harm. The sections that follow define the scope of users covered, the underlying assumptions guiding the framework, and the specific responsibilities required to support safe, lawful, and human-centered adoption of Generative AI.

2.2 SCOPE & APPLICABILITY

These Guidelines apply to a defined set of users who develop, deploy, use, or are affected by Generative AI technologies, with particular attention to applications that can generate or manipulate content (including text, images, audio, video, deepfakes, and lip-sync). The scope of users covered by the Guidelines includes the following categories:

1. Developers and Model Providers

Entities or individuals that design, train, fine-tune, or supply generative AI models, tools, or platforms, including foundation and general-purpose models.

2. Deployers and Service Providers

Organizations that integrate generative AI into products, services, applications, or platforms made available to users, whether in the public or private sector.

3. Institutional and Organizational Users

Public authorities, educational institutions, research bodies, and private organizations that use generative AI systems to support operations, decision-making, communication, or service delivery.

4. Individual Users and Content Creators

Persons who use generative AI tools to create, modify, or distribute content, including students, researchers, professionals, media creators, and developers.

5. High-Impact or Sensitive Use Contexts

Users deploying generative AI in contexts that may significantly affect individuals or society, such as education, media, public information, elections, employment, justice, or content involving real persons.

The Guidelines focus on clarifying **responsibilities proportionate to each user's role and level of control** over generative AI systems. Obligations and safeguards increase with the potential impact and risk of use, ensuring accountability while supporting lawful, innovative, and beneficial applications of generative AI.

2.3 ASSUMPTIONS

These Guidelines are developed based on a set of explicit assumptions that reflect international practice and guidance issued by organizations. These assumptions clarify the context, limits, and intended application of the Guidelines.

1. Generative AI is a rapidly evolving technology

The Guidelines assume that generative AI models, capabilities, and risks will continue to evolve quickly. As a result, the Guidelines are designed to be adaptable and subject to periodic review, rather than static or technology-specific.

2. Generative AI introduces distinct risks beyond traditional AI

It is assumed that generative AI presents unique challenges—such as hallucinations, deepfakes, large-scale misinformation, and training-data opacity—that require targeted policy and governance measures, in addition to general AI principles.

3. Risk and impact vary by context and use case

The Guidelines assume that not all uses of generative AI carry the same level of risk. Policies and governance measures should therefore be **risk-proportionate**, with stronger safeguards applied to high-impact or sensitive uses.

4. Human oversight remains essential

It is assumed that generative AI systems cannot independently ensure accuracy, legality, or ethical compliance. Meaningful human oversight is necessary, especially where outputs may affect individuals' rights, public trust, or critical decisions.

5. Responsibility aligns with control and influence

The Guidelines assume that accountability should be allocated according to the role and level of control exercised by users (developers, deployers, institutional users, or individual users), rather than treating all users equally.

6. Generative AI outputs are probabilistic and may be inaccurate

It is assumed that generative AI systems do not possess understanding or intent and may produce outputs that are misleading or incorrect. Users are therefore responsible for validating outputs before reliance or dissemination.

7. Transparency is a prerequisite for trust

The Guidelines assume that disclosure of AI-generated content, documentation of system limitations, and traceability of outputs are necessary conditions for maintaining trust and enabling accountability.

8. Existing legal and ethical frameworks continue to apply

It is assumed that generative AI does not operate outside existing laws or human-rights obligations. The Guidelines complement, rather than replace, applicable legal, regulatory, and institutional frameworks.

9. Legitimate and beneficial uses must be preserved

The Guidelines assume that generative AI has lawful and socially beneficial applications, including education, research, accessibility, creativity, and innovation. Regulatory approach should therefore mitigate harm without unnecessarily restricting legitimate use.

10. International alignment and interoperability are essential

It is assumed that generative AI is inherently cross-border in development and deployment. Aligning with international principles supports interoperability, legal certainty, and responsible global adoption.

2.4 THE RELEVANCE OF THE GUIDELINES TO DIFFERENT STAKEHOLDERS

Importance of Generative AI Guidelines for Governments

For governments, Generative AI guidelines are essential to protect public trust, fundamental rights, and democratic institutions while enabling responsible innovation. Generative AI systems can influence public information, service delivery, policy design, and civic participation at scale, making ungoverned use a potential risk to transparency, accountability, and the rule of law. Clear guidelines help governments ensure that AI-generated content is transparent, lawful, and subject to human oversight, particularly in sensitive areas such as public communication, education, justice, health and elections. International practice, reflected in the guidance, emphasizes that such guidelines also promote policy coherence, interoperability with global standards, and legal certainty for public-sector adoption.

Importance for Government Institutions and Organizations

For institutions—such as universities, research centers, public bodies, and private organizations—Generative AI guidelines provide a clear policy framework that supports innovation while managing operational, and reputational risks. Institutions increasingly rely on generative AI for content creation, analysis, education, and decision support, yet these systems can produce inaccurate, biased, or misleading outputs. Guidelines clarify acceptable uses, disclosure requirements, data protection responsibilities, and accountability structures, helping institutions avoid misuse, over-reliance, or ethical breaches. By aligning institutional practice with international norms, guidelines also support trust among users, partners, and regulators, and enable consistent adoption across departments and sectors.

Importance for Individuals

For individuals, including students, professionals, creators, and the general public, Generative AI guidelines are important to clarify rights, responsibilities, and expectations. They help individuals understand when and how AI tools may be used responsibly, how to disclose AI assistance, and how to assess the reliability of AI-generated outputs. Guidelines also protect individuals from harm by promoting transparency, safeguarding personal data, and setting limits on deceptive practices such as undisclosed deepfakes or impersonation. In this way, guidelines empower individuals to benefit from generative AI while reducing risks of misuse, dependency, or misinformation.

Importance for Other Stakeholders

For other stakeholders—such as AI developers, service providers, media organizations, civil society, and regulators—Generative AI guidelines establish a shared reference point for responsible conduct and collaboration. They support clearer allocation of responsibility across the AI lifecycle, encourage safety-by-design approaches, and facilitate dialogue between technology providers and society. At the international level, common guidelines help reduce regulatory fragmentation, support cross-border innovation, and promote interoperability. Overall, Generative AI guidelines act as a stabilizing mechanism, ensuring that the rapid adoption of generative technologies delivers societal value while managing risks in a predictable and accountable manner.

2.5 TOP CONCERNS OF GENERATIVE AI MODELS

1. Knowledge Cutoff

Generative AI models rely on training data available up to a fixed point in time and may lack awareness of recent events, updates, or changes. This limitation can lead to outdated or incomplete information if outputs are not independently verified.

2. Bias (Fair/ Unfair)

Generative AI systems may reflect or amplify biases present in training data, leading to unfair or discriminatory outputs affecting individuals or groups. Bias risks require mitigation through data governance, testing, and human oversight.

3. Hallucinations

Generative AI may produce outputs that are fluent and plausible but factually incorrect or entirely fabricated. These "hallucinations" can mislead users if outputs are relied upon without validation.

4. Ethical Concerns (i.e Deepfake Risk, Fake News, Misuse Risk)

Generative AI can be misused to create deceptive or harmful content, including deepfakes and misinformation. Ethical concerns arise when such content undermines trust, harms individuals, or manipulates public opinion.

5. Privacy & Security (Data Breach)

Generative AI systems may pose risks to personal data and information security, including unauthorized data exposure, leakage, or misuse. Strong privacy protections and security safeguards are essential.

6. Lack of Source Verification

Generative AI outputs often do not clearly indicate sources or attribution, making it difficult to verify accuracy, credibility, or originality. Users must independently verify and properly attribute information.

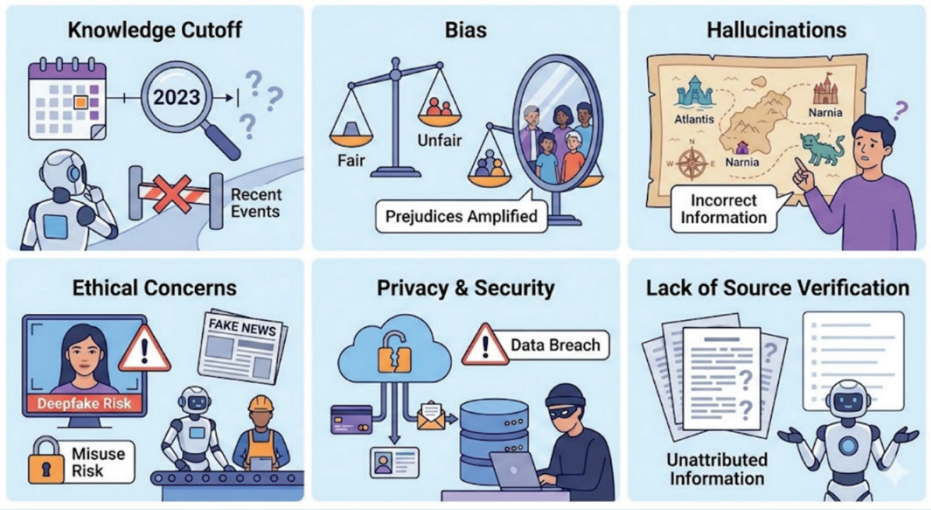

Figure 7: Top Concerns of GEN AI Models

2.6 GUIDELEINES FOR TRUSTWORTHY

GET FAIR AND UNBIASED RESULTS

Fair and unbiased use of Generative AI is essential to ensure that AI-supported outcomes are equitable, reliable, and respectful of all individuals. Because generative AI systems learn from large and diverse datasets, they may reflect or amplify existing social, cultural, or historical biases. Users must therefore apply critical judgment when interacting with these systems, carefully reviewing outputs for unfair assumptions, stereotyping, or discriminatory language. Achieving fair and unbiased results requires neutral input, validation against trusted sources, and human oversight, particularly in contexts that affect people's rights, opportunities, or academic and professional evaluation. By combining responsible use with transparency and accountability, generative AI can support inclusive and ethical outcomes rather than reinforce existing inequities.

You Should:

1. Use **clear, neutral, and precise prompts**, avoiding language that implies stereotypes, preferences, or assumptions about individuals or groups.

2. Review AI-generated outputs **critically**, checking for bias, stereotyping, or unfair representation, especially in content related to gender, ethnicity, religion, age, nationality, or disability.

3. Compare results across **multiple prompts or phrasing styles** to identify inconsistencies or biased patterns in the outputs.

4. Avoid relying on a single AI-generated response for **high-stakes decisions** such as evaluation, grading, hiring, or disciplinary matters.

5. Supplement AI-generated content with **trusted and diverse sources** to balance perspectives and reduce bias.

6. Be aware that Generative AI systems may reflect **biases present in their training data** and that neutrality is not guaranteed.

7. Apply **human judgment and contextual understanding** when interpreting or using AI-generated results.

8. **Report or correct biased outputs** when mechanisms are available, contributing to continuous improvement of the tools.

9. Ensure **that final decisions and conclusions are made by humans**, not solely by AI-generated recommendations.

AVOID HALLUCINATION

Hallucinations in Generative AI refer to situations where an AI system produces outputs that appear fluent, confident, and plausible but are **factually incorrect, misleading, or entirely fabricated**. This phenomenon arises because generative models generate content based on statistical patterns in data rather than verified knowledge or real-world understanding. Hallucinations can occur across text, code, images, audio, and video, and may involve incorrect facts, invented references, false attributions, or inaccurate technical details. When relied upon without verification, hallucinations can lead to errors, misinformation, and inappropriate decisions, particularly in educational, professional, or public-sector contexts. For this reason, international practices emphasizes the need for human oversight, validation against trusted sources, and the responsible use of generative AI outputs rather than treating them as authoritative or definitive.

You should:

1. Treat all Generative AI outputs as probabilistic suggestions, not verified facts or authoritative sources.
2. Verify critical information using reliable, independent sources before relying on or sharing AI-generated content.
3. Avoid using Generative AI as the sole source for factual, legal, medical, technical, or academic information.
4. Request citations or sources where appropriate and independently confirm their accuracy and relevance.
5. Be cautious when Generative AI produces highly confident or detailed responses without clear evidence.
6. Use clear, specific prompts and provide relevant context to reduce ambiguity, which can increase hallucination risk.
7. Break complex tasks into smaller steps and validate outputs incrementally rather than relying on a single response.

8. Test and review AI-generated code, calculations, or technical outputs before use or deployment.

9. Apply human judgment and domain expertise to assess plausibility and consistency of AI-generated results.

10. Avoid over-reliance on Generative AI in high-stakes contexts without human review and approval.

11. When using Generative AI with **private, proprietary, or institutional data**, use **embedding-based retrieval techniques** (such as retrieval-augmented generation (RAG)) to ground model outputs in verified internal sources and reduce the risk of hallucinations.

12. Encourage the model to generate alternative answers or solution approaches, and compare them by outlining their respective advantages, limitations, and underlying assumptions.

13. Support key claims with reputable, authoritative references, and prioritize primary or well-established sources; ensure citations are traceable and verifiable.

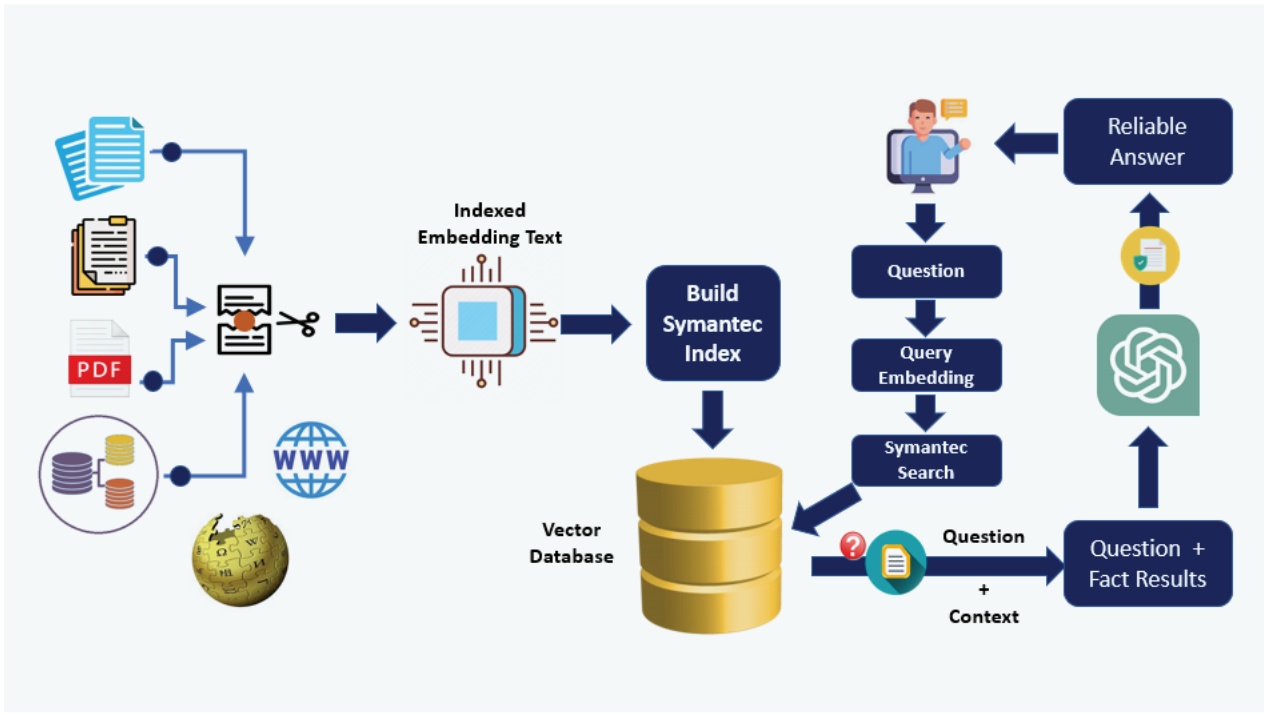

Figure 8: Use of embedding-based retrieval techniques

TARGET RELIABILITY AND SAFETY

To achieve reliable and safe outcomes when using Generative AI, users must recognize that AI-generated outputs are probabilistic and may contain errors or unintended content. Reliability requires careful review, validation, and cross-checking of outputs before they are relied upon or shared. Safety is strengthened by using Generative AI within defined limits, avoiding high-risk or sensitive applications without appropriate oversight. Human judgment and accountability must remain central to interpreting and applying AI-generated results. Clear usage boundaries, transparency, and ethical awareness help reduce misuse and unintended harm. Together, these practices ensure that Generative AI supports informed decision-making while maintaining safety and trust.

You should:

1. Define the system's role and communication style (persona) to ensure the tone, depth, and level of expertise are appropriate for the task and context.

2. Identify the intended audience persona and tailor explanations, terminology, and level of detail to match their background, needs, and familiarity with the subject.

3. Clearly **define the purpose and context** of the task before using Generative AI, including what type of output is needed and how it will be used.

4. Clarify unclear or ambiguous queries by asking targeted follow-up questions to better understand the user's objectives, constraints, and expected outcomes. Reframe complex questions into clearer, more focused components to improve response accuracy and reduce misunderstanding.

5. Use **clear, specific, and unambiguous prompts** to reduce misunderstandings and inaccurate or irrelevant outputs.

6. **Verify and validate all AI-generated outputs**, especially facts, data, code, and technical advice, using trusted and authoritative sources.

7. Treat AI-generated content as **probabilistic and potentially inaccurate**, not as verified or authoritative information.

8. Apply **human judgment and critical thinking** to assess accuracy, relevance, bias, and completeness before relying on or sharing outputs.

9. Avoid using Generative AI as the **sole source of information** for decisions that may affect safety, rights, grades, services, or reputations.

10. Test and review AI-generated code, calculations, or recommendations before deployment or submission.

11. Be alert to **hallucinations**, outdated information, or overconfident language that may mask uncertainty or errors.

12. Use Generative AI tools that provide **transparency about limitations, data usage, and intended purposes**, where available.

13. Escalate to a **human expert or supervisor** when outputs are unclear, high-risk, or may have significant consequences.

TARGET ACCURACY

To get accurate results from Generative AI, users should provide clear, specific, and well-structured inputs that reflect the intended task or question. AI-generated outputs must be critically reviewed and verified, especially when used for academic, technical, or professional purposes. Users should cross-check important information against reliable and authoritative sources rather than rely on a single AI response. Because Generative AI systems may produce incomplete or incorrect content, human judgment remains essential in assessing accuracy and relevance. Refining prompts and asking follow-up questions can help improve output quality, but does not replace validation. Ultimately, responsibility for the correctness and use of the results rests with the user.

You should:

1. Clearly define your question or task and provide specific, well-structured prompts rather than vague instructions.

2. Provide sufficient context, constraints, and background information to guide the Generative AI toward relevant and accurate outputs.

3. Break complex tasks into smaller, manageable steps instead of requesting everything in a single prompt.

4. Verify AI-generated information against reliable and authoritative sources, especially for factual, technical, or academic content.

5. Apply critical thinking and human judgment to review, edit, and correct AI outputs before relying on or submitting them.

6. Test and validate AI-generated code, calculations, or technical recommendations in practice rather than assuming correctness.

7. Avoid relying on Generative AI for information that requires real-time accuracy or up-to-date data unless external verification is performed.

8. Be aware that Generative AI may produce confident but incorrect answers and should be treated as an assistive tool, not an authoritative source.

PROTECT YOUR PRIVACY

To protect your privacy when using Generative AI, users should be cautious about the information they share with AI tools. Personal, sensitive, or confidential data should not be shared unless its use is clearly authorized and necessary. Users should be aware that prompts and outputs may be stored or logged by service providers, which can create privacy risks if information is overshared. AI-generated content should be reviewed to ensure it does not unintentionally reveal private details about individuals or organizations. Where possible, privacy settings and data-minimization options should be enabled. Maintaining privacy ultimately requires informed, responsible use and ongoing human judgment.

You should:

1. Avoid entering or sharing personal or sensitive information (such as national ID numbers, addresses, phone numbers, passwords, images, audio files, video, health or financial data) into Generative AI tools unless explicitly authorized and necessary.

2. Protect other people's privacy by not uploading or generating content that includes personal data about classmates, colleagues, or third parties without their consent.

3. Assume prompts and outputs may be stored or logged by service providers and use Generative AI accordingly.

4. Use anonymized or fictional data when practicing, testing, or learning with Generative AI tools.

5. Follow institutional and legal data-protection rules, including policies on personal data, confidentiality, and acceptable use.

6. Avoid sharing confidential or restricted information, including internal documents, exam materials, proprietary code, or non-public records.

7. Review privacy settings and terms of use of Generative AI tools before using them, especially when creating accounts.

8. Report privacy concerns or suspected data misuse to the relevant institution or authority.

9. Always carefully review the Service Level Agreement (SLA) and the Privacy Policy provided by the Generative AI model vendor before using the service[10].

TARGET TRANSPARENCY AND EXPLAINABILITY

Transparency and Explainability are essential to the responsible use of Generative AI, as they enable users and stakeholders to understand when and how AI systems are influencing outcomes. Generative AI tools should clearly indicate when content is generated or assisted by AI and provide accessible information about their purpose, limitations, and expected behavior. Explainable results support informed decision-making by allowing users to assess the reliability and appropriateness of AI outputs rather than treating them as authoritative. Where Generative AI is used in high-impact or public-facing contexts, additional clarity should be provided to explain how outputs are produced and reviewed. Transparency also supports accountability by making it possible to trace decisions, challenge outcomes, and correct errors. Together, transparency and Explainability help build trust while reducing the risk of misuse, over-reliance, or unintended harm.

You should:

1. Clearly **inform users and audiences** when content, outputs, or recommendations are generated or significantly assisted by Generative AI.

2. Understand and communicate the **intended purpose, limitations, and appropriate use** of the Generative AI system being used.

10 Sample of Privacy Policy: https://privacy.openai.com/policies

3. Avoid presenting AI-generated outputs as fully objective, authoritative, or human-produced without disclosure.

4. Review and assess AI-generated results using **human judgment**, especially when outputs may influence decisions, learning outcomes, or public understanding.

5. Be able to **explain, in plain language**, how Generative AI contributed to the final output, including the role it played in generating, editing, or suggesting content.

6. Maintain **documentation or records** of how Generative AI tools are used in assignments, projects, or institutional processes, where applicable.

7. Verify and, where necessary, **cross-check information** produced by Generative AI against reliable and authoritative sources.

8. Use Generative AI systems that provide **clear information about model capabilities and known limitations**, including risks such as hallucinations or bias.

9. Avoid relying on Generative AI outputs in high-impact contexts unless **meaningful explanation and justification** can be provided.

10. Support transparency measures such as **labelling, disclosure notes, or content provenance mechanisms** when sharing AI-generated content.

TARGET UPTODATE RESULTS

To get the most recent results when using Generative AI, users should be aware that many models rely on training data with a fixed knowledge cutoff and may not reflect the latest developments, events, or updates. Generative AI outputs should -therefore- not be assumed to be current or complete by default. Users are encouraged to cross-check AI-generated information with up-to-date, authoritative sources, especially for time-sensitive topics such as technology, policy, health, or public information. When accuracy and timeliness are critical, Generative AI should be used as a supporting tool, not as the sole source of information. Applying human judgment and verification helps reduce the risk of relying on outdated or inaccurate results.

You should:

1. Be aware that many Generative AI tools have a **knowledge cutoff** and may not reflect the most recent events, regulations, data, or technological developments.

2. **Verify critical or time-sensitive information** using authoritative and current sources such as official websites, government publications, academic journals, or trusted news outlets.

3. Use Generative AI to **draft summaries, or explore ideas**, and then update the output manually with the latest verified information.

4. When available, use Generative AI tools that are **connected to real-time data sources or web search**, and clearly distinguish between AI-generated content and externally verified information.

5. Cross-check dates, figures, legal references, and policy statements before relying on or sharing AI-generated outputs.

6. Avoid using Generative AI as the **sole source** for decisions, publications, or actions that depend on current or rapidly changing information.

7. Clearly note any **assumptions or uncertainties** when AI-generated content may be based on outdated knowledge.

8. Consult subject-matter experts or official guidance when accuracy and timeliness are critical.

KEEP USAGE ETICAL AND AVOIDE MISUSE

Keeping the use of Generative AI ethical and avoiding misuse is essential to maintaining trust, safety, and integrity. Generative AI should be used only for lawful and responsible purposes that respect human rights, dignity, and societal values. Users must not employ these tools to deceive, manipulate, impersonate others, or spread false or harmful information. All AI-generated outputs should be reviewed and verified using human judgment before being relied upon or shared. Personal data, confidential information, and intellectual property must be protected and not misused through AI systems. Ultimately, human responsibility and accountability must remain central to every use of Generative AI.

You should:

1. Use Generative AI only for lawful, legitimate, and constructive purposes that align with ethical values and institutional policies.
2. Avoid using Generative AI to deceive, manipulate, impersonate, or mislead individuals or the public, including through deepfakes or false information.
3. Ensure that AI-generated content is reviewed, verified, and validated before use or dissemination.
4. Clearly disclose the use of Generative AI where transparency is required, especially in academic, professional, or public-facing contexts.
5. Respect human rights, dignity, and cultural values, and avoid generating content that is discriminatory, hateful, or harmful.
6. Protect privacy and personal data by abstaining from sharing sensitive, confidential, or personal information unless authorized and necessary.
7. Refrain from using Generative AI to circumvent rules, safeguards, assessments, or accountability mechanisms.
8. Avoid over-reliance on Generative AI and ensure that human judgment and responsibility remain central to decision-making.
9. Respect intellectual property rights and avoid generating or distributing content that infringes copyright or ownership rights.
10. Report suspected misuse, harmful outputs, or ethical concerns related to Generative AI to the appropriate authority or institution.

2.7 AGENTIC AI GUIDELINES

ONE-LINE PRINCIPLE

When **Generative AI** is used within **Agentic AI** systems, or when an **AI agent is used** within Generative AI tools, system autonomy must always be clearly bounded, auditable, and transparent, and must remain subordinate to human oversight, accountability, and ethical responsibility. Human judgment and control shall remain central to all high-impact decisions and actions undertaken by such systems.

Users should:

- Define **clear objectives, scope, and limits** for agentic AI systems before deployment.
- Ensure Generative AI components operate **only within predefined tasks and permissions**.
- Prohibit open-ended or self-expanding goals without explicit human authorization.

HUMAN OVERSIGHT AND DECISION AUTHORITY

Users should:

- Maintain **meaningful human-in-the-loop or human-in-command oversight**, especially for high-impact actions.
- Require **human approval** for irreversible, high-risk, or externally visible actions (e.g. public communication, system changes, data access).
- Ensure humans retain the **final decision-making authority**, not the agent.

ACCOUNTABILITY AND RESPONSIBILITY

Users should:

- Clearly assign responsibility for agent actions based on **who designs, deploys, configures, and operates** the system.
- Treat outputs and actions generated through agentic workflows as **human-supervised outcomes**, not autonomous decisions.
- Ensure accountability does not shift to the AI system itself.

TRANSPARENCY AND TRACEABILITY

Users should:

- Log agent decisions, prompts, actions, tool calls, and outputs in a traceable and auditable manner.
- Document how Generative AI is used within the agent (e.g. planning, reasoning, content generation).
- Make system limitations and autonomy levels clear to users and stakeholders.

RISK ASSESSMENT AND PROPORTIONAL SAFEGUARDS

Users should:

- **Conduct** risk and impact assessments **before deploying agentic AI systems.**
- **Apply stronger safeguards where agents:**
  - Interact with real-world systems,
  - Affect individuals' rights,
  - Generate or disseminate public-facing content,
  - Operate over longtime horizons.
- Regularly reassess risks as capabilities or contexts change.

SAFETY, SECURITY, AND MISUSE PREVENTION

Users should:

- Implement safeguards against **runaway behavior**, prompt injection, tool misuse, and unintended actions.
- Restrict agent access to sensitive systems, data, or APIs using **least-privilege principles**.
- Include emergency **pause, override, and shutdown mechanisms**.

ACCURACY, RELIABILITY, AND VALIDATION

Users should:

- Recognize that Generative AI outputs within agents are probabilistic and may be incorrect.
- Require validation checks for agent-generated plans, code, or decisions before execution.
- Avoid deploying agentic systems that rely solely on unverified AI-generated reasoning.

DATA PROTECTION AND PRIVACY

Users should:

- Limit the data accessible to agentic systems to what is **strictly necessary**.
- Prevent agents from retaining, reusing, or sharing personal or sensitive data without authorization.
- Ensure compliance with applicable data protection and confidentiality requirements.
- **Ensure that prompts do not unintentionally disclose organizational secrets or sensitive operational details**, recognizing that prompts and outputs may be logged, stored, or reviewed by service providers.

Example:

X Inappropriate prompt (DO NOT use):

"Analyze the attached internal incident report from the Ministry's cybersecurity system and suggest how to fix the vulnerabilities listed, including server IPs, user roles, and access credentials."

- **Why this is a risk:**
  This prompt discloses confidential internal documents, security details, and sensitive operational information, which may be logged, stored, or exposed through the AI service.

✓ **Appropriate prompt (SAFE alternative):**

"Provide general best practices for improving cybersecurity incident response processes in public-sector organizations, without using or referencing any real internal systems or data."

ETHICAL USE AND HUMAN-CENTRED DESIGN

Users should:

- Design agentic AI to **support and augment human capabilities**, not replace human judgment or responsibility.
- Avoid agent behaviors that deceive, manipulate, or unduly influence users.
- Consider societal and ethical impacts, particularly in public-sector, educational, or sensitive contexts.

MONITORING, REVIEWING, AND CONTINUOUS IMPROVEMENT

Users should:

- Continuously monitor agent behavior and outcomes in real-world operation.
- Establish procedures for incident reporting, correction, and learning.
- Periodically review agent design, permissions, and use cases in line with evolving international guidance.

2.8 DEEPFAKE GUIDELINES

ONE-LINE GUIDING PRINCIPLE

Deepfake and lip-sync technologies may be used for legitimate purposes only where transparency, consent, accountability, and safeguards against deception and harm are ensured.

TRANSPARENCY AND DISCLOSURE

- Users and providers **must clearly disclose** when audio, video, images, or text are generated or materially altered using generative AI, including lip-sync technologies.
- Disclosure should be provided through **visible labels**, notices, or **technical provenance mechanisms** such as metadata or watermarking.
- Removing, concealing, or falsifying such disclosure should be prohibited.

PREVENTION OF DECEPTION AND MISREPRESENTATION

- Deepfake and lip-sync technologies **must not be used to mislead audiences** into believing that a real person said or did something they did not.
- Particular care must be taken where content appears realistic or authoritative, as realism increases the risk of deception.

PROTECTION OF INDIVIDUALS AND CONSENT

- The **non-consensual use** of a person's likeness, voice, or facial movements—including realistic lip-sync—is prohibited or strictly restricted.
- Special protection must apply to **minors**, vulnerable individuals, and.

HIGH-RISK CONTEXT RESTRICTIONS

- Use of deepfakes and lip-sync should be **restricted or subject to enhanced safeguards** in sensitive contexts, including:
  - political communication and elections
  - public information and government messaging
  - journalism and news media
  - education and training environments
- In such contexts, additional transparency, review, and approval mechanisms should apply.

ACCOUNTABILITY AND RESPONSIBILITY

- Responsibility for deepfake and lip-sync content lies with those who **create, deploy, or distribute** the content, not with the AI system itself.
- Organizations should maintain **clear accountability chains** and internal approval processes for high-risk uses.

LEGITIMATE AND BENEFICIAL USES

- Legitimate uses of deepfake and lip-sync technologies—such as **film production, dubbing and accessibility, education, satire, and artistic expression**—may be permitted, provided that:
  - transparency is maintained,
  - consent is obtained where applicable,
  - and the content does not cause harm or deception.

RISK ASSESSMENT AND PROPORTIONAL SAFEGUARDS

- Before deploying deepfake or lip-sync technologies, users should conduct **risk and impact assessments** focusing on potential harm, deception, and societal impact.
- Safeguards should be **proportionate to the realism, scale, and audience reach** of the content.

MONITORING, REPORTING, AND REMEDIATION

- Providers and platforms should implement mechanisms to:
  - monitor misuse,
  - enable rapid reporting of harmful or deceptive content,
  - and support **timely correction, labelling, or removal** where harm occurs.

ETHICAL AND HUMAN-CENTRED USE

- Deepfake and lip-sync technologies should be designed and used to **support legitimate expression and creativity**, not to manipulate, coerce, or undermine trust.
- Human dignity, autonomy, and public trust must remain central considerations.

2.9 GENERATIVE AI IN EDUCATION AND SCIENTIFIC RESEARCH

In education and scientific research, generative AI may be used as a supportive tool to enhance learning and research activities when human judgment, originality, and academic responsibility are maintained. It can assist with literature review, data analysis, language improvement, and exploration of ideas, but its use must remain transparent and appropriately disclosed. All AI-generated outputs should be critically reviewed, verified, and validated by students or researchers before use or publication.

Generative AI must not replace independent thinking, misrepresent authorship, fabricate or falsify data, manipulate results, or bypass assessment, peer-review, or ethical approval processes. The production of misleading, plagiarized, or unverified scientific content is not permitted. International practice reflects these expectations, as leading publishers and journals have adopted explicit policies on generative AI use and disclosure. These approaches align with the **UNESCO guidelines**.

Permissible Usage

- Language Enhancement: Authors may use AI tools to improve the readability and language of their manuscripts. However, this must be done under human oversight, with authors taking full responsibility for the content.
- Research Design and Methods: AI tools can be employed as part of the research design or methods (e.g., AI-assisted imaging). Such use must be described in a reproducible manner in the methods section, including details about the AI tool used.
- Disclosure Requirement: Any use of generative AI or AI-assisted technologies in the writing process must be transparently disclosed in the manuscript. A statement will appear in the published work to inform readers.

Non-Permissible Usage

- Authorship Attribution: AI tools cannot be credited as authors or co-authors. Authorship implies responsibilities that can only be attributed to and performed by humans.
- Content Generation: Using AI to generate scientific, pedagogic, or medical insights, draw scientific conclusions, or provide clinical recommendations is prohibited.
- Image Creation or Alteration: The use of generative AI or AI-assisted tools to create or alter images in submitted manuscripts is not permitted, except when part of the research design or methods, as previously noted.
- Undisclosed AI Use: Failing to disclose the use of AI tools in the manuscript is against Elsevier's policies and may be considered a breach of publishing ethics.

2.10 DISCLOSURE

Key principle

Disclosure promotes trust and integrity; it does not excuse misuse or reduce human responsibility.

Why disclosure is required?

Disclosure of Generative AI use is essential to maintain transparency, academic integrity, and accountability in education, research, and professional work. Because generative AI systems can produce fluent but uncertain or derivative outputs, disclosure allows instructors, reviewers, publishers, and stakeholders to properly evaluate originality, responsibility, and reliability.

International practice, reflected in policies issued by major academic publishers and bodies such as the Committee on Publication Ethics (COPE) and guidance from UNESCO, consistently requires clear disclosure while affirming that responsibility remains with the human author.

When is disclosure required?

Users should disclose the use of Generative AI when it has been used to:

- Generate or substantially modify text, code, images, data analysis, or figures
- Assist in drafting, summarizing, translating, or restructuring content
- Support research activities beyond basic proofreading or spelling checks
- Contribute to outputs submitted for assessment, publication, or public use

Generative AI must not be listed as an author or co-author, and disclosure does not transfer responsibility away from the human user.

How to disclose?

Disclosure should be clear, concise, and specific, indicating:

- The name of the Generative AI tool used
- The purpose for which it was used
- The extent of its contribution
- Confirmation that the user reviewed and validated the outputs

Disclosure may appear in an acknowledgements section, methods section, footnote, appendix, or dedicated AI-use statement, depending on institutional or publisher requirements.

GENERATIVE AI DISCLOSURE TEMPLATE

Template – Short Disclosure Statement

Generative AI Use Disclosure:

This work used [name of Generative AI tool/system] for the purpose of [e.g. language editing, code assistance, idea generation]. All outputs were reviewed, verified, and edited by the author(s), who remain fully responsible for the content, accuracy, and integrity of this work.

Template – Academic / Research Disclosure (Extended)

Use of Generative Artificial Intelligence:

Generative AI tools were used in the preparation of this work. Specifically, [tool name and version] was used to assist with [describe the task, e.g. summarizing background literature, improving language clarity, generating draft code]. The AI system did not generate original research findings or make substantive intellectual contributions. All AI-assisted content was critically reviewed, validated, and revised by the author(s), who take full responsibility for the final content.

Template – Student Assignment Disclosure

AI Assistance Declaration:

I used [name of Generative AI tool] to support [specific task]. I confirm that this submission reflects my own understanding and work, that AI-generated outputs were reviewed and verified, and that I remain fully responsible for the content submitted.

ANNEX 1:

EFFECTIVE PROMPTING FOR GENERATIVE AI

PROMPT ENGINEERING PATTERNS

The effectiveness of a conversational LLM's responses is strongly influenced by the clarity and quality of the user's prompts.

Prompt Engineering is identified as the skillful crafting, utilization of prompts to interact with various systems and tools effectively.

- **Prompt:** A prompt can be described as an instruction or signal that initiates an action or response. In the context of AI conversations, a prompt is usually a line of a text or a question that guides the AI model's response.
- **Types of Prompts:** Text prompts are most common, but voice-based prompts are increasingly prevalent with advancements in speech recognition technology. Image prompts are also used in certain applications, such as image captioning tasks.

The Power of Prompt Engineering:

- **Why Prompt Engineering Matters:** Effective prompt engineering can significantly enhance efficiency, accuracy, and user experience in various applications.
- **Applications in Different Fields:** Prompt engineering is utilized in chatbots for customer service, digital artworks, virtual assistants like Siri and Alexa, and data analysis tools, specialized applications in Healthcare, Education, and etc…

GENERAL TIPS FOR DESIGNING PROMPTS

1. **Clear Instructions:** Give precise commands to the AI about the task, such as "Write", "Classify", "Summarize", "Translate", "Order", etc.
2. **Provide Examples and Steps:** Offer samples and a step-by-step process to guide the AI's understanding and execution.
3. **Iterative Process**: Recognize that prompting is not an exact science and requires experimentation and refinement. Try different prompts, styles, and approaches to achieve the desired outcome.
4. **Interact with the AI:** Engage in a dialogue with the AI, providing feedback and requesting modifications to improve results.
5. **Incorporate Human Expertise:** Include your knowledge and insights into the prompts to enhance the AI's output.
6. **Precision:** Enhance the precision of a prompt to get accurate responses and ensure comprehensive coverage of relevant information.
7. **To do or not to do?** Avoid saying what not to do but say what to do instead.
8. **Ethical Considerations:** Discuss ethical considerations related to prompt engineering, such as bias in language models and the potential impact of prompts on user behavior.
9. **Understanding Context:** Crafting prompts that consider the context of the task or query, including relevant keywords and language nuances.

Example of bad and good prompts:

X Bad prompt

"Write about generative AI guidelines."

✓ Fully qualified prompt

You are an expert in AI governance and public-sector digital policy.

Draft a concise guideline (600–800 words) for university staff and students on the responsible use of Generative AI in education and research.

The guideline must:

- Cover: fairness and bias, hallucinations, privacy & data protection, disclosure of AI use, and academic integrity.
- Clearly distinguish between permitted uses (e.g. language polishing, idea exploration) and not permitted uses (e.g. undisclosed authorship, fabricated data, deepfake misuse).
- Use clear headings and bullet points where appropriate.
- Use neutral, non-technical language suitable for a mixed audience (students, lecturers, admin staff).
- End with a short checklist (5–7 items) that a student can follow before submitting any AI-assisted work.

PROMPT ENGINEERING TECHNIQUES

Prompt engineering techniques focus on designing and structuring inputs that guide large language models (LLMs) toward accurate, relevant, and controllable outputs. Key practices include clear task framing, where the goal and constraints are stated explicitly; context enrichment, adding background details or examples to reduce ambiguity; and instruction hierarchy, sequencing steps so the model follows logical reasoning.

Techniques such as few-shot prompting (showing sample responses), chain-of-thought prompting (encouraging step-by-step reasoning), and role prompting (assigning the model a persona or expertise) can significantly improve response quality. Iterative refinement—testing, evaluating, and adjusting prompts—is essential, as prompt quality strongly influences output performance across tasks such as reasoning, summarization, and creativity.

Prompt engineering is now **widely known and practiced** in AI work—especially among developers, analysts, researchers, and advanced users. It has proven helpful for improving accuracy, reducing ambiguity, and directing large-language-model behavior. However, it is not fully deterministic: **the same prompt can produce slightly different outputs**, and results still depend on the underlying model, data quality, and task complexity. So it is best viewed as a **practical skill and methodology**, not a perfect or universal solution. Professionals typically combine prompt engineering with testing, validation, and human review—especially in critical fields.

1. ZERO SHOT PROMPT

Definition

Zero-shot prompting means asking the AI to perform a task without providing any example outputs. The model relies only on the task description, context, and constraints specified by the user. This works best when the instructions are clear, structured, and unambiguous.

Example

- Translate the sentence 'La vita è bella' to English.
- Summarize the following paragraph: Data governance is **non-negotiable** for lawful, safe, and fair AI. Partial compliance directly increases legal, ethical, and operational risk, especially for medium- and high-impact systems.

2. ONE SHOT PROMPT

Definition

One-shot prompting means asking the AI to perform a task after providing one example output that demonstrates the desired structure, tone, or style. The model uses this single example as a reference to better understand the expected response format while still relying mainly on your task description and constraints.

Example

- Question: "What is the capital of the United States?"
  Answer: "The capital of the United States is Washington D.C."
- Question: "What is the capital of France?"
  Answer:

3. N-SHOT PROMPT

Definition

N-shot prompting means asking the AI to perform a task after providing multiple example outputs (N examples) that demonstrate the desired structure, tone, and reasoning style. By learning from several examples, the model can better generalize patterns and produce more consistent, accurate responses—especially for complex or domain-specific tasks.

Example:

Prompt 1 — Input:

"Generative AI may produce outdated information, so users should verify key facts before relying on them."

Output (Style Reference):

"AI-generated information is not always current. Users should check important details against trusted and up-to-date sources before using or sharing them."

Prompt 2 — Input:

"AI tools can sometimes reflect biases found in their training data."

Output (Style Reference):

"AI systems may unintentionally mirror existing social or data-driven biases. Human oversight is important to ensure fair and responsible use."

Prompt 3: Explain the concept of hallucinations in Generative AI using the same neutral, educational tone as the example outputs.

4. Chain of thought

Definition

Chain-of-thought prompting means asking the AI to reason step by step before giving a final answer. The prompt explicitly requests intermediate reasoning, such as listing assumptions, breaking the problem into parts, or following a sequence of logical steps. This helps the model produce more structured, transparent, and reliable outputs for complex tasks, especially in analysis, planning, and problem-solving.

Example Prompt

Prompt: Analyze the main risks of using Generative AI in university student assessments and propose practical mitigation measures, using step-by-step reasoning.

Output format:

- Section 1: Step-by-step reasoning (identify risks, then map each to controls)
- Section 2: Final concise list of 5–7 mitigation recommendations in bullet points

Constraints:

- Make each reasoning step clear and logically ordered (no long paragraphs)
- Avoid technical jargon and legal references unless strictly necessary
- Do not include speculative or sensational risks; focus on realistic, documented concerns.

5. The Reflection technique

Definition

When you provide an answer, please explain the reasoning and assumptions behind your selection of software frameworks. If possible, use specific examples or evidence with associated code samples to support your answer of why the framework is the best selection for the task. Moreover, please address any potential ambiguities or limitations in your answer, in order to provide a more complete and accurate response.

Example Prompt

1. Reflection step

- Whenever you generate an answer:

1. Explain the reasoning and key assumptions behind the selection
2. Provide at least one short code snippet or concrete usage example
3. Identify potential ambiguities, trade-offs, or limitations in the recommendation
4. Briefly mention 1–2 alternative frameworks and why they were not chosen

Recommend the most suitable web framework for building a secure, large-scale public-service portal (e.g. e-government platform), and then reflect on the reasoning.

6. The Persona and audience persona Prompt

Definition

Persona prompting means assigning the AI a specific professional role or identity (e.g., policy advisor, cybersecurity analyst, university lecturer) so that responses match the expected tone, expertise level, and perspective.

Audience-persona prompting means clearly defining who the response is intended for (e.g., students, senior leaders, developers), so the language, depth, and style are appropriate for that group. Using both together helps produce clearer, more relevant, and context-aware outputs.

Example Prompt

Prompt: Explain the risks of deepfake misuse in simple, responsible, and informative language.

Persona (AI role): Public-sector AI governance advisor

Audience persona: First-year university students with general digital-literacy knowledge

7. Fact check List pattern

Definition

The Fact-Check List Pattern means explicitly asking the AI to verify key claims or statements against reliable sources and present the verification in a structured checklist format. Each claim is listed with its status (e.g., accurate / partially accurate / fake statistics / misleading / false), followed by a short explanation and a citation to an authoritative source. This technique helps reduce misinformation, especially in policy, education, and public-facing contexts where accuracy and transparency are essential.

Example Prompt

- If I ask you a question, compile a list of the key facts. Insert this fact list at the end of the summary.
- **Question:** Review and fact-check statements about Generative AI risks and provide clear verification.

8. The Cognitive Verifier technique

Definition

When the user asks a question, the system is expected to generate three additional questions that would help the user give a more accurate answer. The system should assume that the user knows little about the topic that both the system and user are discussing and that it should define any terms that are not general knowledge. When the user answers the three questions, the system combines the answers to produce the final answers to the original question.

Example Prompt

- When you are asked a question,
- Generate additional 5 questions to understand my context. Combine my answers to try to identify my weakness, and recommend a training course for me.
- When I ask about the impact of climate change on agriculture,
- **Question:** Generate three more questions to clarify aspects of the changes I ask about. Once these are answered, integrate the information in a response to the initial query.

9. Alternative answers pattern

Definition

Alternative Answers Pattern (also called **Alternative Approaches Pattern**) in prompt engineering is a structured technique used to get an AI model (like a large language model) to generate multiple distinct responses or solutions to the same problem or task, instead of a single answer. It encourages the model to explore **different ways of thinking** or **different strategies** for solving the same question, helping you compare options and pick the best one.

Example Prompt

- For every task I give you, list the best 3 alternative approaches with their pros and cons.
- Question: How can I secure my file on my private computer?

10. Provide reputable authorized reference

Definition

A "reputable authorized reference" is a **source of information that is widely recognized as trustworthy, credible, and authoritative in its field**, and that can be relied on to support factual claims or conclusions in research, writing, or decision-making. It typically comes from **experts, established institutions, peer-reviewed publications, official organizations, or respected media outlets** whose accuracy and reliability are verifiable and acknowledged by others in the domain.

Example Prompt

- Act as an expert in Agile software development. Explain how to run a daily Scrum meeting, including its purpose, structure, participants, time-boxing, and common best practices.
- **Question:** For any factual statements, cite reputable and authoritative sources and include the direct source links in your response.

11. Question refinement

Definition

Question Refinement (often called the **Question Refinement Pattern**) is the practice of **iteratively improving or clarifying a question or query** so that a generative AI model understands your intent better and can produce a more accurate, relevant, and focused answer. It typically involves examining an initial question, revising it to reduce ambiguity or add missing context, and then using the improved version to guide the AI's response.

Example Prompt

- Whenever I ask a question about the software development, suggest a better version of the question that considers the Pros and Cons of each selection, like performance, security, ease of development.
- I need to develop a mechanism to update customer profile after the customer place an order on the website, which approach is suitable? Fire a database trigger or run another SQL update statement to be run immediately after the customer place his order.

12. Flipped interaction pattern

Definition

The Flipped Interaction Pattern is a structured interaction technique where the **AI model leads the dialogue by asking the user questions** to gather the necessary information, instead of the typical approach where the user asks and the model answers. This "flip" in roles helps the model uncover context or missing details it needs to fulfill a task more effectively.

Example Prompt

- From now on, I would like you to ask me questions to deploy a .NET application to IIS on AWS. When you have enough information to deploy the application, create a script to automate the deployment.
- Ask me the questions one by one. I prefer MCQ questions
- Ask questions until I make less than three mistakes in a row. Ask me the first question.
- I would like you to ask me questions to test my knowledge of the Solar System.
- You should ask questions until I answer all of them correctly.

ANNEX 2:
LIST OF ACRONYMS AND ABBREVIATIONS

CONCEPTS AND TECHNOLOGIES

- **AI** – Artificial Intelligence
- **GenAI** – Generative Artificial Intelligence
- **GPAI** – General-Purpose Artificial Intelligence
- **LLM** – Large Language Model
- **RAG** – Retrieval-Augmented Generation
- **GAN** – Generative Adversarial Network
- **API** – Application Programming Interface
- **SLA** – Service Level Agreement
- **Deepfake** – AI-generated or AI-manipulated synthetic media that imitates real persons or events
- **Lip-sync** – AI-based technique that synchronizes mouth movements with synthetic or altered audio

ORGANIZATIONS

- **OECD** – Organization for Economic Co-operation and Development
- **UNESCO** – United Nations Educational, Scientific and Cultural Organization
- **G7** – Group of Seven
- **GPAI (Initiative)** – Global Partnership on Artificial Intelligence
- **ITU** – International Telecommunication Union
- **ISO** – International Organization for Standardization
- **IEC** – International Electrotechnical Commission
- **IEEE** – Institute of Electrical and Electronics Engineers
- **NIST** – National Institute of Standards and Technology (United States)
- **COPE** – Committee on Publication Ethics
- **EU** – European Union
- **CAC** – Cyberspace Administration of China
- **SDAIA** – Saudi Data and Artificial Intelligence Authority

ANNEX 3:
REFERENCES

1. EU AI Act Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance.
2. OECD updates AI Principles to stay abreast of rapid technological developments OECD updates AI Principles to stay abreast of rapid technological developments | OECD
3. OpenAI — Prompt engineering best practices: https://platform.openai.com/docs/guides/prompt-engineering
4. Microsoft Learn — Prompt engineering techniques: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering
5. Stanford HAI — Understanding prompt design for LLMs: https://hai.stanford.edu/news/promise-and-pitfalls-prompting-large-language-models
6. UNESCO — Guidance for Generative AI in Education and Research https://www.unesco.org/en/generative-ai
7. Evolving with innovation: The 2024 OECD AI Principles update https://oecd.ai/en/wonk/evolving-with-innovation-the-2024-oecd-ai-principles-update
8. Updated OECD AI Principles to keep up with novel and increased risks from general purpose and generative AI private-ai.com
9. Guidelines for the use of generative AI in the public administration https://oecd.ai/en/dashboards/policy-initiatives/guidelines-for-the-use-of-generative-ai-in-the-public-administration-6317?utm_source=chatgpt.com
10. Sample of Privacy Policy: https://privacy.openai.com/policies
11. Commission publishes first draft of Code of Practice on marking and labelling of AI-generated content https://digital-strategy.ec.europa.eu/en/news/commission-publishes-first-draft-code-practice-marking-and-labelling-ai-generated-content
12. Overview of Draft Measures on Generative AI https://www.chinalawtranslate.com/en/overview-of-draft-measures-on-generative-ai/
13. Singapore's Model AI Governance Framework for Generative AI Singapore's Model AI Governance Framework for Generative AI : Clyde & Co
14. Rigorous AI research to enable advanced AI governance https://www.aisi.gov.uk/
15. G7 Summit – Hiroshima Process documents (official) Leaders_Communique_01_en.pdf