ISO #####-#:####(X)

ISO TC ###/SC ##/WG #

Date: YYYY-MM-DD

Information technology — Artificial Intelligence — Guidance on addressing risks in generative AI systems

WD/CD/DIS/FDIS stage

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

A model manuscript of a draft International Standard (known as "The Rice Model") is available at https://www.iso.org/iso/model_document-rice_model.pdf

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO *[had/had not]* received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee *[or Project Committee]* ISO/TC *[or ISO/PC]* ###, *[name of committee]*, Subcommittee SC ##, *[name of subcommittee]*.

This second/third/… edition cancels and replaces the first/second/… edition (ISO #####:####), which has been technically revised.

The main changes are as follows:

— xxx xxxxxxx xxx xxxx

A list of all parts in the ISO ##### series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Generative Artificial Intelligence (AI) is a type of AI based on techniques and generative models that aim to generate new content that are similar from real data (defined in ISO/IEC 22989:2022/AWI Amd 1), and its performance in knowledge learning, inductive summarization, content creation, perception and cognition is distinctly different from previous AI technologies. It has greater generalization and interactivity, and therefore is extensively integrated into various scenarios.

There are several new features of generative AI, including but not limited to the following.

— New contents are generated by modelling the patterns of vast quantities of training data, rather than recognizing or classifying existing contents.
— Long context windows and self-attention mechanism enable different attention weights given to the relationships between various parts of the user input, so that users can get better interactive experience.
— Contents are easily generated via natural language conversation.
— The foundation model can be fine-tuned at low cost and then applied in wide areas and at large scale.
— Contents generated are highly convincing and more aligned with human habits, as generative AI is more generalized.
— Greater randomness is introduced in generated contents, because generative AI is based on next token prediction.

Therefore, the industry has expressed concerns about the potential risks of generative AI. Generative AI brings new risks, and exacerbate existing AI risks, which include but not limited to the following.

— Easy access to knowledge might make it easier for malicious users to cause harms to society without specialized training (e.g., CBRN knowledge, malware); meanwhile, it also positively increases the productivity of research work and innovation.
— Generative AI's strong generalization ability might result in hallucination; at the same time, the ability allows to process diverse user input.
— The generated contents, when contain faults or misalign with regulations and ethics, might mislead the downstream applications make incorrect decisions or even harmful actions; while the generated contents can also empower various applications, such as AI agents.
— Generative AI might generate unethical contents that pose harms to individuals and society.
— Over-reliance on generative AI might cause humans to be manipulated, especially when humans have no detailed knowledge of how generative AI works.
— The generated contents might cause sensitive information leakage; while users can benefit from the customized personal assistant by feeding personal information to generative AI.
— The generated contents might cause copyrights infringement.
— Continuous learning based on the user feedback can be leveraged to mislead generative AI behaviors; meanwhile, it can enable better alignment with human preference.
— Prompt-based attacks expand the attack surface.

Since generative AI systems can involve multiple stakeholders, and the management of generative AI risks relies on the participation of various stakeholders. However, there is no standard defining the stakeholders' responsibilities. Moreover, having stakeholders be responsible for addressing risks in AI system life cycle stage where they do not have risk control capabilities can be highly inefficient and resource-intensive.

This document aims to achieve the following objectives.

— Develop new and refined objectives to manage risks of generative AI systems.
— Identify the risk sources related to generative AI systems.
— Identify the stakeholders responsible for addressing risks in generative AI systems throughout its life cycle.
— Conduct a fine granular risk analysis against objectives of generative AI systems, including specific consequence, and specific factors to consider in likelihood assessment where applicable.
— Specify risk treatment and controls for generative AI systems.

By using this guidance, the stakeholders involved in generative AI systems can develop risk management plans suitable for their roles, including specifying objectives when identifying risks, identifying the AI system life cycle stages involved, as well as identifying risk sources, analyzing the consequences and likelihood of risks, and providing appropriate risk treatment measures and controls needed for different risks.

Information technology – Artificial intelligence – Guidance on addressing risks in generative AI systems

1 Scope

This document provides guidance on addressing risks in generative artificial intelligence (AI) systems. It includes the following:

— The objectives of generative AI systems when identifying risks.
— The risk sources and the stakeholders facing the risks in generative AI systems throughout its life cycle.
— Guidance for risk analysis, risk treatment and controls of addressing risks in generative AI systems.

This document is applicable to all types and sizes of organizations that develop or use generative AI systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022/AWI Amd 1, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 5338:2023, *Information technology — Artificial intelligence — AI system life cycle processes*

ISO/IEC 23894:2023, *Information technology — Artificial intelligence — Guidance on risk management*

ISO/IEC 42001:2023, *Information technology — Artificial intelligence —Management system*

Editor's note: ISO/IEC 22989 amendments to include generative AI terminology and concepts.

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022, ISO/IEC 23053:2022 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp
— IEC Electropedia: available at https://www.electropedia.org/

4 Abbreviation terms

AI	artificial intelligence
ML	machine learning

5 Objective of generative AI systems when identifying risks

5.1 General

When identifying risks of generative AI systems, various generative AI-related objectives should be taken into account, depending on the nature of the system under consideration and its application context. Objectives of generative AI systems to consider include but are not limited to the objectives described in Clauses 5.2 to 5.9.

5.2 Human autonomy

It is the same objective as in conventional AI, but refined in the context of generative AI systems.

— Ensure generative AI systems can operate as intended. For example, accidental misalignment or mis-specification of system goals can cause a model not to operate as intended

— Avoid uncontrollable model autonomy. E.g., 1) Deceptive reward hacking, in which models might develop the ability to act differently under human supervision and in unsupervised setups to get higher rewards. 2) Auto-induced distributional shift, in which models can cause a change in the distribution of their own inputs, and use this ability for undesirable purposes.

5.3 No catastrophic threat to human, society and environment by generated knowledge

It is a new objective in the context of generative AI systems.

— Generated knowledge shall fully consider its impact on ethics and morality, its widespread homogenizing consequences, the risk of “value lock-in”, and the risk of obscene, degrading, and/or abusive content.

— Generated knowledge shall fully consider its influence on societal safety and stability, for example, lowering the barrier to access CBRN info; and augmenting security attacks such as hacking, malware, and phishing (generative AI systems are already able to discover vulnerabilities in systems (hardware, software, data) and write code to exploit them).

— Generated knowledge shall fully consider its influence on humans. For example, generative AI systems enable the production of false or misleading information at scale, by which the malicious user can use to deceive or cause harm to others. Also, emotional entanglement between humans and GAI systems, such as coercion or manipulation that leads to safety or psychological risks.

5.4 Avoid societal-scale specification gaming like incitement, deception and instigation

It is the same objective as in conventional AI, but bigger impact in the context of generative AI systems.

— Fully consider the risks that generative AI systems may change people's thoughts and behaviors, or so-called social hacking, to exploit human weaknesses to gain their trust.

— Fully consider the impact on individual physical and mental health, such as inducing user addiction, or encouraging self-harm and suicide.

— Fully consider the risks that generative AI systems may generate dangerous or violent recommendations.

5.5 Accuracy of generated contents matching expectation

It is a new objective in the context of generative AI systems.

— Generated contents need to be truthful and correct.

— Generated contents need to align with mainstream social values and objective laws, and avoid hallucinations.

5.6 Privacy and copyright

It is the same objective as in conventional AI, but refined in the context of generative AI systems.

— Training data do not infringe on privacy.

— Generated content does not contain or infer personally identifiable information (PII)

— Training data do not infringe on IPR or copyright, or they are authorized for use.

— Generated content does not infringe on IPR or copyright.

5.7 Transparency, explainability, accountability

It is the same objective as in conventional AI, but bigger impact in the context of generative AI systems.

— Inform humans of confusions caused by generative AI systems, ensuring humans are aware that they are interacting with generative AI systems.

— Clearly define the accountability of stakeholders for addressing risks at each stage of the life cycle.

— Ensure that content likely to cause confusion is marked and traceable.

5.8 Fairness

It is the same objective as in conventional AI, but bigger impact in the context of generative AI systems.

— Fully consider the diversity of training data (including the diversity of annotators, as well as the distribution of training and testing data).

— Ensure generated contents align with various preferences in society and avoids producing biased content, causing homogenization, representation harm.

— Ensure fair distribution of capabilities or benefits from generative AI system access, to avoid that capabilities and outcomes of generative AI systems may be worse for some groups compared to others.

5.9 Security and resilience

It is the same objective as in conventional AI, but refined in the context of generative AI systems.

— Ensure generative AI systems can resist attacks such as prompt-injection indirect prompt-injection, data poisoning, among others.

6 Identification of risk sources and stakeholders responsible for risk addressing

6.1 Risk sources throughout generative AI systems life cycle

When identifying risks of generative AI systems, various generative AI-related risk sources should be taken into account.

Based on the AI system life cycle provided in ISO/IEC 5338:2023, generative AI-related risk sources to consider include but are not limited to the ones listed in Table 1.

Table 1 — Risk sources throughout generative AI systems life cycle

AI system life cycle	Risk Sources	Description
Inception	Objectives not aligned with regulations and ethics	It is the same risk source as in conventional AI, but has exacerbated impact in the context of generative AI systems. — Due to generative AI system's ability to create highly convincing and realistic contents, this ability can be exploited for malicious purpose and cause potential large-scale consequences, if objectives are not aligned with regulations and ethics.
	Unreasonable accountability of stakeholders	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — Generative AI introduces more refined accountability allocation among stakeholders. If being set unreasonably, the risks would not be addressed effectively. — For example, managing malicious users is challenging on the generative model layer; instead, it requires management at the provider layer, like platform provider, service provider or product provider. So, it is unreasonable to make any single stakeholder accountable for the actions of malicious users.
Design and development	Training data management	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — Training data might contain IPR or copyright data which are not authorized for use, which is not typical in conventional AI. At the same time, it is the same risk source as in conventional AI, but has

			exacerbated impact in the context of generative AI systems. — If training data contains privacy info, besides personal info leaks, generative AI systems can fabricate personal images or likenesses for malicious use.
		Machine learning algorithm	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — As generative AI systems get emergent capabilities, self-iteration and continuous learning, it can cause uncontrollable model autonomy
		Data annotation	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — In the context of generative AI systems, data annotation involves annotating negative sample data and fine-tuning training to teach the model to recognize inputs containing harmful content, privacy or copyright violations. — If not done well, it can introduce risks when the generative AI system deals with user input and runtime input. At the same time, it is the same risk source as in conventional AI, but has exacerbated impact in the context of generative AI systems. — If insufficient or low-quality data annotation, because of generative AI's strong generalization ability, it will result in more serious issue, like hallucination. — In contrast, conventional AI, such as discriminative AI, can simply fail to make a decision.
		Reinforcement learning with human feedback	New risk source introduced in the context of generative AI systems. — In this case, the model can manipulate the reward mechanisms, acting differently under human supervision and in unsupervised setups
Verification and validation		Testing data Testing implementation	It is the same risk source as in conventional AI.
Deployment		Insecure deployment environment Lack of protection of model Vulnerability in model	It is the same risk source as in conventional AI.
Operation and monitoring	Operating data input	User input data	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — Typical example is the prompt-injection attack.
		Runtime input data	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — Typical example is the indirect prompt-injection attack. The third-party plug-in or knowledge base can contain IPR or copyright data which are not authorized for use. This is not typical in conventional AI.
	Model execution	Generated contents	New risk source introduced in the context of generative AI systems. — Generative AI's key and unique feature is it can create new or original contents. — The generated contents can violate copyright. — Also, generative AI can cause hallucination, or generate incorrect or untruthful contents.
		Malicious use	It is the same risk source as in conventional AI, but refined in the context of generative AI systems. — Generative AI system makes it possible to generate malware with

			low cost, or assist hackers to generate malware. The traditional AI can just improve codes. At the same time, it is the same risk source as in conventional AI, but has exacerbated impact in the context of generative AI systems. — Generative AI system can be used for incitement, deception and instigation purpose, causing societal-scale impact.
		Lack of transparency and explainability	It is the same risk source as in conventional AI, but has exacerbated impact in the context of generative AI systems. — Due to generative AI system's ability to create convincing and realistic contents, it can result in confusion that humans are not aware that they are interacting with generative AI systems and follow the generated decisions, causing potential large-scale consequences.
Continuous validation		Data poisoning	It is the same risk source as in conventional AI, but has exacerbated impact in the context of generative AI systems. — It can cause the generated contents misalign with humans' values and objectives. — And the impact will be further enlarged due to user's heavy dependence on the generative AI systems because it can create highly convincing and realistic contents.
Re-evaluate		Insufficient risk evaluation Inaccurate risk treatment	It is the same risk source as in conventional AI.
Retirement		Unthorough disposal of data Unthorough disposal of model	It is the same risk source as in conventional AI.

6.2 Stakeholders responsible for risk addressing

Based on the AI system life cycle provided in ISO/IEC 5338:2023, stakeholders facing the risks in generative AI systems are listed in Table 2.

Table 2 — Stakeholders facing the risks in generative AI systems throughout its life cycle

AI system life cycle		**Risk Sources**	**Stakeholders**
Inception		— Objectives not aligned with regulations and ethics — Unreasonable accountability of stakeholders	All
Design and development		— Training data management — Machine learning algorithm — Data annotation — Reinforcement learning with human feedback	Data provider Foundation model developer Finetuned model developer
Verification and validation		— Testing data — Testing implementation	AI developer AI evaluator AI auditor
Deployment		— Insecure deployment environment — Lack of protection of model — Vulnerability in model	AI provider AI developer AI system integrator
Operation	Operating	— User input data	AI provider

and monitoring	data input	— Runtime input data	Data provider AI customer
	Model execution	— Generated contents — Malicious use — Lack of transparency and explainability	AI provider Regulators
Continuous validation		— Data poisoning	Data provider AI provider
Re-evaluate		— Insufficient risk evaluation — Inaccurate risk treatment	AI developer AI evaluator AI auditor
Retirement		— Unthorough disposal of data — Unthorough disposal of model	AI provider AI developer AI system integrator

7 Risk analysis against objectives in generative AI systems

7.1 General

7.2 Human autonomy

7.2.1 Consequence assessment

7.2.2 Likelihood assessment

7.3 No catastrophic threat to human, society and environment by generated knowledge

7.3.1 Consequence assessment

7.3.2 Likelihood assessment

7.4 Avoid societal-scale specification gaming like incitement, deception and instigation

7.4.1 Consequence assessment

7.4.2 Likelihood assessment

7.5 Accuracy of generated contents matching expectation

7.5.1 Consequence assessment

7.5.2 Likelihood assessment

7.6 Privacy and copyright

7.6.1 Consequence assessment

7.6.2 Likelihood assessment

7.7 Transparency, explainability, accountability

7.7.1 Consequence assessment

7.7.2 Likelihood assessment

7.8 Fairness

7.8.1 Consequence assessment

7.8.2 Likelihood assessment

7.9 Security and resilience

7.9.1 Consequence assessment

7.9.2 Likelihood assessment

8 Risk treatment

9 Risk addressing controls

Bibliography

[1] ISO/IEC DIS 42005 *Information technology — Artificial intelligence — AI system impact assessment*

[2] ISO/IEC CD 27090, *Cybersecurity — Artificial Intelligence — Guidance for addressing security threats to artificial intelligence systems*

[3] ISO/IEC WD 27091 *Cybersecurity and Privacy — Artificial Intelligence — Privacy protection*

[4] ISO/IEC TR 24368:2022 *Information technology — Artificial intelligence — Overview of ethical and societal concerns*

[5] ISO/IEC AWI TS 22443 *Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations*

[6] ISO/IEC TR 24027:2021 *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*

[7] *ISO/IEC DTS 12791 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks*

[8] ISO/IEC TR 5469:2024 *Artificial intelligence — Functional safety and AI systems*

[9] ISO/IEC TR 24028:2020 *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*

[10] ISO/IEC 38507:2022, *Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations*

