ISO Form 4
NEW WORK ITEM PROPOSAL (NP)

Circulation date: 2024-10-29 **Closing date for voting:** 2025-01-22	**Reference number:** ISO/IEC NP TS 25568 ISO/IEC JTC 1/SC 42 **N 1950**
Proposer ISO/IEC JTC 1/SC 42	
Secretariat ANSI	

A proposal for a new work item within the scope of an existing committee shall be submitted to the secretariat of that committee.
A proposal for a new project committee shall be submitted to the Central Secretariat, which will process the proposal in accordance with ISO/IEC Directives, Part 1, Clause 2.3.
Guidelines for proposing and justifying new work items or new fields of technical activity (Project Committee) are given in ISO/IEC Directives, Part 1, Annex C.
IMPORTANT NOTE: Proposals without adequate justification and supporting information risk rejection or referral to the originator.

☒ The proposer confirms that this proposal has been drafted in compliance with Annex C of ISO/IEC Directives, Part 1.

PROPOSAL

(to be completed by the proposer, following discussion with committee leadership if appropriate)

TITLE

English title:

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

French title:

(In the case of an amendment, revision or a new part of an existing document, show the reference number and current title)

SCOPE

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

PURPOSE AND JUSTIFICATION

Generative Artificial Intelligence (AI) is a type of AI based on techniques and generative models that aim to generate new content (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.
- The automatic generation of contents (texts, pictures, sounds) has a strong impact for the work organization in many professions where such contents have been so far created by human beings.

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.

While several existing ISO standards are also dealing with AI risks, there are still some gaps.

- For ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management, it covers risk management process, AI-related objectives and AI risk sources, but it does not provide objectives related to generative AI system, does not provide risk sources related to generative AI systems, and it lacks the granularity of risk analysis against objectives of generative AI systems, including what are the specific risk consequence, and what are the specific factors to consider in likelihood assessment.
- For ISO/IEC 42001:2023 Information technology — Artificial intelligence —Management system, it covers controls of addressing risk related to the design and operation of AI systems, but it does not provide controls related to the risks of generative AI systems.
- Also, the existing ISO standards fail to identify the stakeholders responsible for addressing risks related to generative AI systems.

Therefore, this standard 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.

Sustainable Development Goals (SDGs)

Goal 3: Good Health and Well-Being for People
Goal 4: Quality Education
Goal 5: Gender Equality
Goal 8: Decent Work and Economic Growth
Goal 9: Industry, Innovation, and Infrastructure
Goal 10: Reducing Inequalities
Goal 11: Sustainable Cities and Communities
Goal 12: Responsible Consumption and Production
Goal 16: Peace, Justice and Strong Institutions

Preparatory work

☐ A draft is attached ☒ An outline is attached ☐ An existing document serving as the initial basis is attached

The proposer is prepared to undertake the preparatory workrequired:

☒ Yes ☐ No

If a draft is attached to this proposal:

Please select from one of the following options:

☒ The draft document can be registered at Preparatory stage (WD – stage 20.00)

☐ The draft document can be registered at Committee stage (CD – stage 30.00)

☐ The draft document can be registered at enquiry stage (DIS – stage 40.00)

If the attached document is copyrighted or includes copyrighted content:

☐ The proposer confirms that copyright permission has been granted for ISO to use this content in compliance with the ISO/IEC Directives, Part 1 (see also the Declaration on copyright).

Is this proposal for an ISO management System Standard (MSS)?

☐ Yes ☒ No

Note: If yes, this proposal must have an accompanying justification study. Please see the Consolidated Supplement to the ISO/IEC Directives, Part 1, Annex SL or Annex JG

Indication of the preferred type to be developed

☐ International Standard ☒ Technical Specification

☐ Publicly Available Specification *

* While a formal NP ballot is not required to start developing a PAS (no eForm04), the NP form may provide useful information for the committee P-members to consider when deciding to initiate a Publicly Available Specification.

Proposed Standard Development Track (SDT – to be discussed by the proposer with the committee manager or ISO/CS)

☐ 18 months ☒ 24 months ☐ 36 months

Draft project plan (as discussed with committee leadership)

Proposed date for first meeting: 2025-01-27

Dates for key milestones: Circulation of 1st Working Draft (if any) to experts: 2025-01-27

Committee Draft consultation (if any): 2025-08-20

DIS submission*:

Publication*: 2027-01-20

* Target Dates for DIS submission and Publication should be set a few weeks ahead of the limit dates automatically determined when selecting the SDT.

NOTE: ISO/Meetings and ISO/Projects allow you to register and continuously update the meeting dates and project target dates during the development of the project.

Known patented items (see ISO/IEC Directives, Part 1 for important guidance)

☐ Yes ☒ No

If "Yes", provide full information as annex

Co-ordination of work: To the best of your knowledge, has this or a similar proposal been submitted to another standards development organization?

☐ Yes ☒ No

If “Yes”, please specify which one(s):

Listing of relevant documents (such as standards and regulations) at international, regional and national level

Listing of relevant documents (such as standards and regulations) at international, regional and national level (Please see ISO/IEC Directives, Part 1, Annex C, Clause C.4.6)
Detailed gap analysis on existing related ISO standards.
1. ISO/IEC 22989:2022/AWI Amd 1 Information technology - Artificial intelligence – concepts and terminology
Introduction: it defines terminology and concepts of AI, including AI stakeholders, AI system life cycle, AI ecosystem, applications of AI system, etc. It provides standardized terminology and AI

concepts to better understand and use AI among different stakeholders.
Gap analysis: this proposal will refer the terminology and concepts in ISO/IEC 22989:2022, especially the term of generative AI for consistency.

2. ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes
Introduction: it defines AI system life cycle, including AI system life cycle processes, AI system life cycle model and its different stages of inception, design and development, verification and validation, deployment, operation and monitoring, continuous validation, re-evaluation and retirement.
Gap analysis: based on the AI system life cycle model in ISO/IEC 5338:2023, this proposal will describe the risk sources in generative AI systems throughout life cycle.

3. ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management
Introduction: it provides guidance for organization using AI to manage AI risks, including principles, framework, and process.
Gap analysis: based on the AI risk management process defined in ISO/IEC 23894:2023, this proposal will develop the objectives when identifying risks, fine granular risk analysis against objectives, risk treatment, and controls related to addressing risks in generative AI systems.

4. ISO/IEC 42001:2023 Information technology — Artificial intelligence —Management system
Introduction: it specifies the requirements and provides guidance for establishing, implementing, maintaining and continually improving an AI management system, including the controls of addressing risk related to the design and operation of AI systems.
Gap analysis: this proposal will further provide the specific controls of addressing risks in generative AI systems.

5. ISO/IEC DIS 42005 Information technology — Artificial intelligence — AI system impact assessment
Introduction: it provides guidance for organizations performing AI system impact assessments for individuals and societies that can be affected by an AI system and its foreseeable applications. It includes how this AI system impact assessment process can be integrated into an organization’s AI risk management and AI management system.
Gap analysis: based on the guidance of implementing an AI system impact assessment process in ISO/IEC DIS 42005, this proposal will describe the impact and consequence of risks related to generative AI systems.

6. ISO/IEC CD 27090 Cybersecurity — Artificial Intelligence — Guidance for addressing security threats to artificial intelligence systems
Introduction: it provides guidance to address security threats in AI systems, including how to detect and mitigate such threats.
Gap analysis: for the risks and controls related to security threats in generative AI systems, this proposal will refer the work of ISO/IEC CD 27090.

7. ISO/IEC WD 27091 Cybersecurity and Privacy — Artificial Intelligence — Privacy protection
Introduction: it provides guidance for organizations to address privacy risks in AI systems and ML models, including privacy risks identification, consequences evaluation and treatment.
Gap analysis: for the risks and controls related to privacy risks in generative AI systems, this proposal will refer the work of ISO/IEC WD 27091.

8. ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concerns
Introduction: it provides overview of AI ethical and societal concerns, including themes and principles, and examples of practices for building and using ethically and societally acceptable AI.
Gap analysis: this proposal will refer the themes and principles in the area of AI ethical and societal concerns in ISO/IEC TR 24368:2022.

9. ISO/IEC AWI TS 22443 Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations
Introduction: it provides guidance on how an organization can identify and address societal concerns and ethical considerations during the life cycle of AI systems that can potentially harm individuals and society.
Gap analysis: for the risks and controls related to societal concerns and ethical considerations in generative AI systems, this proposal will refer the work of ISO/IEC AWI TS 22443.

10. ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
Introduction: it describes the types of bias in AI systems, and methods for assessing and addressing bias.
Gap analysis: for the risks and addressing methods related to bias in AI systems, if applicable for generative AI systems, this proposal will refer the work of ISO/IEC TR 24027:2021.

11. ISO/IEC DTS 12791 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks

Introduction: it describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. It provides how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks.

Gap analysis: for the risks and addressing methods related to unwanted bias in classification and regression machine learning tasks, if applicable for generative AI systems, this proposal will refer the work of ISO/IEC DTS 12791.

12. ISO/IEC TR 5469:2024 Artificial intelligence — Functional safety and AI systems

Introduction: it defines the relationship between functional safety and AI systems, including the functional safety risk factors and mitigation measures in AI systems.

Gap analysis: for the risks and mitigation measures related to functional safety in AI systems, if applicable for generative AI systems, this proposal will refer the work of ISO/IEC TR 5469:2024.

13. ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

Introduction: it provides approaches and measures to building and improving the trustworthiness of AI systems, including application of risk management, stakeholders, vulnerabilities and threats of AI systems, and mitigation measures.

Gap analysis: for the application of risk management, stakeholders, vulnerabilities and threats of AI systems, and mitigation measures, if applicable for generative AI systems, this proposal will refer the work of ISO/IEC TR 24028:2020.

14. ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

Introduction: it provides guidance for the governing body of an organization to govern the use of AI, including the polices on governance of decision-making, governance of data use, culture and values, compliance, and risk.

Gap analysis: the guidance for the governing body of an organization to the policies on risk in ISO/IEC 38507:2022 applies in this proposal.

15. ISO/IEC 25059:2023 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems

Introduction: it defines the quality model for AI systems, including AI system quality in use model containing societal and ethical risk mitigation.

Gap analysis: this proposal has different scope with ISO/IEC 25059:2023.

16. ISO/IEC 25053:2022 Framework for artificial intelligence (AI) systems using machine learning (ML)

Introduction: it defines the ML framework (including data, models, tasks and other elements). It provides unified ML terminology and framework for different stakeholders.

Gap analysis: this proposal will refer the ML terminology in ISO/IEC 25053:2022 to ensure consistency.

Identification and description of relevant affected stakeholder categories (Please see ISO CONNECT)

	Benefits/Impacts/Examples
Industry and commerce - large industry	Large industry can be the providers developing and offering generative AI systems, or can be the providers offering products, systems and services that utilize generative AI systems. The large industry can use this standard to understand their role throughout generative AI system life cycle, and implement appropriate controls provided by the standard to address risks in generative AI systems throughout its life cycle. This is helpful to increase the public trust and market acceptance in the use of generative AI systems provided by large industry.
Industry and commerce - SMEs	SMEs can be the AI partner, such as generative AI system integrators, providers of data used by generative AI systems, providers of evaluation service for generative AI systems, and providers for audit service to assess conformance of generative AI systems to standards and regulatory requirements. In addition to the benefits and impacts listed for large industry, SMEs will benefit from this standard, introducing new market opportunities of providing services related to

	addressing risks of generative AI systems.
Government	Government will benefit from the guidance, allowing them to analyze, identify, manage and monitor risks in generative AI systems, either for internal or for external (and public) use.
Consumers	Consumers can enhance understanding of the objectives of generative AI systems, and the role and accountability of stakeholders of generative AI systems, therefore increase trust into generative AI systems.
Labour	This standard can be used to analyze the risks of generative AI systems on labour market, including positive aspect like creating new opportunities and negative aspect like making existing jobs redundant.
Academic and research bodies	This standard can drive artificial intelligence innovation by helping researchers to develop new methods that are both cutting-edge and risk-controllably sound.
Standards application businesses	
Non-governmental organizations	Increase trust in human autonomy, accuracy of generated contents and fairness of generative AI systems, and therefore increase the acceptance and applicability of generative AI systems for a variety of NGOs, such as environment monitoring and protection, advocation for the right the education, etc.
Other (please specify)	

Liaisons: A listing of relevant external international organizations or internal parties (other ISO and/or IEC committees) to be engaged as liaisons in the development of the deliverable.	**Joint/parallel work:** **Possible joint/parallel work with:** □ IEC (please specify committee ID) □ CEN (please specify committee ID) □ Other (please specify)

A listing of relevant countries which are not already P-members of the committee.

Note: The Committee Manager shall distribute this NP to the ISO members of the countries listed above to ask if they wish to participate in this work

Proposed Project Leader (name and e-mail address) Qing An anqing.aq@alibaba-inc.com	**Name of the Proposer** (include contact information) Heather Benko hbenko@ansi.org

This proposal will be developed by:

☒ An existing Working Group: ISO/IEC JTC 1/SC 42/WG 3 Trustworthiness

□ A new Working Group:

(Note: establishment of a new Working Group requires approval by the parent committee)

□ The TC/SC directly

□ To be determined:

Supplementary information relating to the proposal

☒ This proposal relates to a new ISO document

□ This proposal relates to the adoption as an active project of an item currently registered as a Preliminary Work Item

□ This proposal relates to the re-establishment of a cancelled project as an active project

Other:

Maintenance agencies (MA) and registration authorities (RA)

□ This proposal requires the designation of a maintenance agency. If so, please identify the potential candidate:

□ This proposal requires the designation of a registration authority. If so, please identify the potential candidate:

NOTE: Selection and appointment of the MA or RA are subject to the procedure outlined in ISO/IEC Directives, Part 1, Annex G and Annex H.

☒ Annex(es) are included with this proposal (provide details)

Additional information/question(s)