INTERNATIONAL STANDARD

ISO/IEC 23894

First edition
2023-02

Information technology — Artificial intelligence — Guidance on risk management

Technologies de l'information — Intelligence artificielle — Recommandations relatives au management du risque

Reference number
ISO/IEC 23894:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

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 document 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 or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see https://patents.iec.ch).

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. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

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 and www.iec.ch/national-committees.

Introduction

The purpose of risk management is the creation and protection of value. It improves performance, encourages innovation and supports the achievement of objectives.

This document is intended to be used in connection with ISO 31000:2018. Whenever this document extends the guidance given in ISO 31000:2018, an appropriate reference to the clauses of ISO 31000:2018 is made followed by AI-specific guidance, if applicable. To make the relationship between this document and ISO 31000:2018 more explicit, the clause structure of ISO 31000:2018 is mirrored in this document and amended by sub-clauses if needed.

This document is divided into three main parts:

Clause 4: Principles – This clause describes the underlying principles of risk management. The use of AI requires specific considerations with regard to some of these principles as described in ISO 31000:2018, Clause 4.

Clause 5: Framework – The purpose of the risk management framework is to assist the organization in integrating risk management into significant activities and functions. Aspects specific to the development, provisioning or offering, or use of AI systems are described in ISO 31000:2018, Clause 5.

Clause 6: Processes – Risk management processes involve the systematic application of policies, procedures and practices to the activities of communicating and consulting, establishing the context, and assessing, treating, monitoring, reviewing, recording and reporting risk. A specialization of such processes to AI is described in ISO 31000:2018, Clause 6.

Common AI-related objectives and risk sources are provided in Annex A and Annex B. Annex C provides an example mapping between the risk management processes and an AI system life cycle.

Information technology — Artificial intelligence — Guidance on risk management

1 Scope

This document provides guidance on how organizations that develop, produce, deploy or use products, systems and services that utilize artificial intelligence (AI) can manage risk specifically related to AI. The guidance also aims to assist organizations to integrate risk management into their AI-related activities and functions. It moreover describes processes for the effective implementation and integration of AI risk management.

The application of this guidance can be customized to any organization and its context.

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 31000:2018, *Risk management — Guidelines*

ISO Guide 73:2009, *Risk management — Vocabulary*

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

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 31000:2018, ISO/IEC 22989:2022 and ISO Guide 73:2009 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 Principles of AI risk management

Risk management should address the needs of the organization using an integrated, structured and comprehensive approach. Guiding principles allow an organization to identify priorities and make decisions on how to manage the effects of uncertainty on its objectives. These principles apply to all organizational levels and objectives, whether strategic or operational.

Systems and processes usually deploy a combination of various technologies and functionalities in various environments, for specific use cases. Risk management should take into account the whole system, with all its technologies and functionalities, and its impact on the environment and stakeholders.

AI systems can introduce new or emergent risks for an organization, with positive or negative consequences on objectives, or changes in the likelihood of existing risks. They also can necessitate

specific consideration by the organization. Additional guidance for the risk management principles, framework and processes an organization can implement is provided by this document.

NOTE Different International Standards have significantly different definitions of the word "risk." In ISO 31000:2018 and related International Standards, "risk" involves a negative or positive deviation from the objectives. In some other International Standards, "risk" involves potential negative outcomes only, for example, safety-related concerns. This difference in focus can often cause confusion when trying to understand and properly implement a conformant risk management process.

ISO 31000:2018, Clause 4 defines several generic principles for risk management. In addition to guidance in ISO 31000:2018, Clause 4, Table 1 provides further guidance on how to apply such principles where necessary.

Table 1 — Risk management principles applied to artificial intelligence

	Principle	Description (as given in ISO 31000:2018, Clause 4)	Implications for the development and use of AI
a)	Integrated	Risk management is an integral part of all organizational activities.	No specific guidance beyond ISO 31000:2018.
b)	Structured and comprehensive	A structured and comprehensive approach to risk management contributes to consistent and comparable results.	No specific guidance beyond ISO 31000:2018.
c)	Customized	The risk management framework and process are customized and proportionate to the organization's external and internal context related to its objectives.	No specific guidance beyond ISO 31000:2018.

Table 1 *(continued)*

	Principle	Description (as given in ISO 31000:2018, Clause 4)	Implications for the development and use of AI
d)	Inclusive	Appropriate and timely involvement of stakeholders enables their knowledge, views and perceptions to be considered. This results in improved awareness and informed risk management.	Because of the potentially far-reaching impacts of AI to stakeholders, it is important that organizations seek dialog with diverse internal and external groups, both to communicate harms and benefits, and to incorporate feedback and awareness into the risk management process. Organizations should also be aware that the use of AI systems can introduce additional stakeholders. The areas in which the knowledge, views and perceptions of stakeholders are of benefit include but are not restricted to: — Machine learning (ML) in particular often relies on the set of data appropriate to fulfil its objectives. Stakeholders can help in the identification of risks regarding the data collection, the processing operations, the source and type of data, and the use of the data for particular situations or where the data subjects can be outliers. — The complexity of AI technologies creates challenges related to transparency and explainability of AI systems. The diversity of AI technologies further drives these challenges due to characteristics such as multiple types of data modalities, AI model topologies, and transparency and reporting mechanisms that should be selected per stakeholders' needs. Stakeholders can help to identify the goals and describe the means for enhancing transparency and explainability of AI systems. In certain cases, these goals and means can be generalized across the use case and different stakeholders involved. In other cases, stakeholder segmentation of transparency frameworks and reporting mechanisms can be tailored to relevant personas (e.g. "regulators", "business owners", "model risk evaluators") per the use case. — Using AI systems for automated decision-making can directly affect internal and external stakeholders. Such stakeholders can provide their views and perceptions concerning, for example, where human oversight can be needed. Stakeholders can help in defining fairness criteria and also help to identify what constitutes bias in the working of the AI system.

Table 1 *(continued)*

	Principle	Description (as given in ISO 31000:2018, Clause 4)	Implications for the development and use of AI
e)	Dynamic	Risks can emerge, change or disappear as an organization's external and internal context changes. Risk management anticipates, detects, acknowledges and responds to those changes and events in an appropriate and timely manner.	To implement the guidance provided by ISO 31000:2018, organizations should establish organizational structures and measures to identify issues and opportunities related to emerging risks, trends, technologies, uses and actors related to AI systems. Dynamic risk management is particularly important for AI systems because: — The nature of AI systems is itself dynamic, due to continuous learning, refining, evaluating, and validating. Additionally, some AI systems have the ability to adapt and optimize based on this loop, creating dynamic changes on their own. — Customer expectations around AI systems are high and can potentially change quickly as the systems themselves do. — Legal and regulatory requirements related to AI are frequently changing and being updated. Integration with the management systems on quality, environmental footprints, safety, healthcare, legal or corporate responsibility, or any combination of these maintained by the organization, can also be considered to further understand and manage AI-related risks to the organization, individuals and societies.
f)	Best available information	The inputs to risk management are based on historical and current information, as well as on future expectations. Risk management explicitly takes into account any limitations and uncertainties associated with such information and expectations. Information should be timely, clear and available to relevant stakeholders.	Taking into account the expectation that AI affects the way individuals interact with and react to technology, it is advisable for organizations engaged in the development of AI systems to keep track of relevant information available regarding the further uses of the AI systems that they developed, while users of AI systems can maintain records of the uses of those systems throughout the entire lifetime of the AI system. As AI is an emerging technology and constantly evolving, historical information can be limited, and future expectations can change quickly. Organizations should take this into account. The internal use of AI systems should be considered, if any. Tracking the use of AI systems by customers and external users can be limited by intellectual property, contractual or market-specific restrictions. Such restrictions should be captured in the AI risk management process and updated when business conditions warrant revisiting.

Table 1 *(continued)*

	Principle	Description (as given in ISO 31000:2018, Clause 4)	Implications for the development and use of AI
g)	Human and cultural factors	Human behaviour and culture significantly influence all aspects of risk management at each level and stage.	Organizations engaged in the design, development or deployment of AI systems, or any combination of these, should monitor the human and cultural landscape in which they are situated. Organizations should focus on identifying how AI systems or components interact with pre-existing societal patterns that can lead to impacts on equitable outcomes, privacy, freedom of expression, fairness, safety, security, employment, the environment, and human rights broadly.
h)	Continual improvement	Risk management is continually improved through learning and experience.	The identification of previously unknown risks related to the use of AI systems should be considered in the continual improvement process. Organizations engaged in the design, development or deployment of AI systems or system components, or any combination of these, should monitor the AI ecosystem for Performance successes, shortcomings and lessons learned, and maintain awareness of new AI research findings and techniques (opportunities for improvement).

5 Framework

5.1 General

The purpose of the risk management framework is to assist the organization in integrating risk management into significant activities and functions. The guidance provided in ISO 31000:2018, 5.1 applies.

Risk management involves assembling relevant information for an organization to make decisions and address risk. While the governing body defines the overall risk appetite and organizational objectives, it delegates the decision-making process of identifying, assessing and treating risk to management within the organization.

ISO/IEC 38507[1] describes additional governance considerations for the organization regarding the development, purchase or use of an AI system. Such considerations include new opportunities, potential changes to the risk appetite as well as new governance policies to ensure the responsible use of AI by the organization. It can be used in combination with the risk management processes described in this document to help guide the dynamic and iterative organizational integration described in ISO 31000:2018, 5.2.

5.2 Leadership and commitment

The guidance provided in ISO 31000:2018, 5.2 applies.

In addition to the guidance provided in ISO 31000:2018, 5.2 the following applies:

Due to the particular importance of trust and accountability related to the development and use of AI, top management should consider how policies and statements related to AI risks and risk management are communicated to stakeholders. Demonstrating this level of leadership and commitment can be critical for ensuring that stakeholders have confidence that AI is being developed and used responsibly.

The organization should therefore consider issuing statements related to its commitment to AI risk management to increase confidence of their stakeholders on their use of AI.

Top management should also be aware of the specialized resources that can be needed to manage AI risk, and allocate those resources appropriately.

5.3 Integration

The guidance provided in ISO 31000:2018, 5.3 applies.

5.4 Design

5.4.1 Understanding the organization and its context

The guidance provided in ISO 31000:2018, 5.4.1 applies.

In addition to guidance provided in ISO 31000:2018, 5.4.1, Table 2 lists additional factors to consider when understanding the external context of an organization.

Table 2 — Consideration when establishing the external context of an organization

Generic guidance provided by ISO 31000:2018, 5.4.1 Organizations should consider at least the following elements of their external context:	**Additional guidance for organizations engaged in AI** Organizations should additionally consider, but not exclusively, the following elements:
— The social, cultural, political, legal, regulatory, financial, technological, economic and environmental factors, whether international, national, regional or local;	— Relevant legal requirements, including those specifically relating to AI. — Guidelines on ethical use and design of AI and automated systems issued by government-related groups, regulators, standardization bodies, civil society, academia and industry associations. — Domain-specific guidelines and frameworks related to AI.
— Key drivers and trends affecting the objectives of the organization;	— Technology trends and advancements in the various areas of AI. — Societal and political implications of the deployment of AI systems, including guidance from social sciences.
— External stakeholders' relationships, perceptions, values, needs and expectations;	— Stakeholder perceptions, which can be affected by issues such as lack of transparency (also referred to as opaqueness) of AI systems or biased AI systems. — Stakeholder expectations on the availability of specific AI-based solutions and the means by which the AI models are made available (e.g. through a user interface, software development kit).
— Contractual relationships and commitments;	— How the use of AI, especially AI systems using continuous learning, can affect the ability of the organization to meet contractual obligations and guarantees. Consequently, organizations should carefully consider the scope of relevant contracts. — Contractual relationships during the design and production of AI systems and services. For example, ownership and usage rights of test and training data should be considered when provided by third parties.
— The complexity of networks and dependencies;	— The use of AI can increase the complexity of networks and dependencies.

Table 2 *(continued)*

Generic guidance provided by ISO 31000:2018, 5.4.1 Organizations should consider at least the following elements of their external context:	**Additional guidance for organizations engaged in AI** Organizations should additionally consider, but not exclusively, the following elements:
— (guidance beyond ISO 31000:2018).	— An AI system can replace an existing system and, in such a case, an assessment of the risk benefits and risk transfers of an AI system versus the existing system can be undertaken, considering safety, environmental, social, technical and financial issues associated with the implementation of the AI system.

In addition to guidance provided in ISO 31000:2018, 5.4.1, Table 3 lists additional factors to consider when understanding the internal context of an organization.

Table 3 — Consideration when establishing the internal context of an organization

Generic guidance provided by ISO 31000:2018, 5.4.1 Organizations should consider at least the following elements of their internal context:	**Additional guidance for organizations engaged in AI** Organizations should additionally consider, but not exclusively, the following elements:
— Vision, mission and values;	— No specific guidance beyond ISO 31000:2018
— Governance, organizational structure, roles and accountabilities;	— No specific guidance beyond ISO 31000:2018
— Strategy, objectives and policies;	— No specific guidance beyond ISO 31000:2018
— The organization's culture;	— The effect that an AI system can have on the organization's culture by shifting and introducing new responsibilities, roles and tasks.
— Standards, guidelines and models adopted by the organization;	— Any additional international, regional, national and local standards and guidelines that are imposed by the use of AI systems.
— Capabilities, understood in terms of resources and knowledge (e.g. capital, time, people, intellectual property, processes, systems and technologies);	— The additional risks to organizational knowledge related to transparency and explainability of AI systems. — The use of AI systems can result in changes to the number of human resources needed to realize a certain capability, or in a variation of the type of resources needed, for instance, deskilling or loss of expertise where human decision-making is increasingly supported by AI systems. — The specific knowledge in AI technologies and data science required to develop and use AI systems. — The availability of AI tools, platforms and libraries can enable the development of AI systems without there being a full understanding of the technology, its limitations and potential pitfalls. — The potential for AI to raise issues and opportunities related to intellectual property for specific AI systems. Organizations should consider their own intellectual property in this area and ways that intellectual property can affect transparency, security and the ability to collaborate with stakeholders, to determine whether any steps should be taken.

Table 3 *(continued)*

Generic guidance provided by ISO 31000:2018, 5.4.1 Organizations should consider at least the following elements of their internal context:	**Additional guidance for organizations engaged in AI** Organizations should additionally consider, but not exclusively, the following elements:
— Data, information systems and information flows;	— AI systems can be used to automate, optimize and enhance data handling. — As consumers of data, additional quality and completeness constraints on data and information can be imposed by AI systems.
— Relationships with internal stakeholders, taking into account their perceptions and values;	— Stakeholder perception, which can be affected by issues such as lack of transparency of AI systems or biased AI systems. — Stakeholder needs and expectations can be satisfied to a greater extent by specific AI systems. — The need for stakeholders to be educated on capabilities, failure modes and failure management of AI systems.
— Contractual relationships and commitments;	— Stakeholder perception, which can be affected by different challenges associated with AI systems such as potential lack of transparency and unfairness. — Stakeholder needs and expectations can be satisfied by specific AI systems. — The need for stakeholders to be educated on capabilities, failure modes and failure management of AI systems. — Stakeholders' expectations of privacy, and individual and collective fundamental rights and freedoms.
— Interdependencies and interconnections;	— The use of AI systems can increase the complexity of interdependencies and interconnections.

In addition to the guidance provided in ISO 31000:2018, 5.4.1, organizations should consider that the use of AI systems can increase the need for specialized training.

5.4.2 Articulating risk management commitment

The guidance provided in ISO 31000:2018, 5.4.2 applies.

5.4.3 Assigning organizational roles, authorities, responsibilities and accountabilities

The guidance provided in ISO 31000:2018, 5.4.3 applies.

In addition to the guidance of ISO 31000:2018, 5.4.3, top management and oversight bodies, where applicable, should allocate resources and identify individuals:

— with authority to address AI risks;

— with responsibility for establishing and monitoring processes to address AI risks.

5.4.4 Allocating resources

The guidance provided in ISO 31000:2018, 5.4.4 applies.

5.4.5 Establishing communication and consultation

The guidance provided in ISO 31000:2018, 5.4.5 applies.

5.5 Implementation

The guidance provided in ISO 31000:2018, 5.5 applies.

5.6 Evaluation

The guidance provided in ISO 31000:2018, 5.6 applies.

5.7 Improvement

5.7.1 Adapting

The guidance provided in ISO 31000:2018, 5.7.1 applies.

5.7.2 Continually improving

The guidance provided in ISO 31000:2018, 5.7.2 applies.

6 Risk management process

6.1 General

The guidance provided in ISO 31000:2018, 6.1 applies.

Organizations should implement a risk-based approach to identifying, assessing, and understanding the AI risks to which they are exposed and take appropriate treatment measures according to the level of risk. The success of the overall AI risk management process of an organization relies on the identification, establishment and the successful implementation of narrowly scoped risk management processes on strategic, operational, programme and project levels. Due to concerns related but not limited to the potential complexity, lack of transparency and unpredictability of some AI-based technologies, particular consideration should be given to risk management processes at the AI system project level. These system project level processes should be aligned with the organization's objectives and should be both informed by and inform other levels of risk management. For example, escalations and lessons learned at the AI project level should be incorporated at the higher levels, such as the strategic, operational and programme levels, and others as applicable.

The scope, context and criteria of a project-level risk management process are directly affected by the stages of the AI system's life cycle that are in the scope of the project. Annex C shows possible relations between a project-level risk management process and an AI system life cycle (as defined in ISO/IEC 22989:2022).

6.2 Communication and consultation

The guidance provided in ISO 31000:2018, 6.2 applies.

The set of stakeholders that can be affected by AI systems can be larger than initially foreseen, can include otherwise unconsidered external stakeholders and can extend to other parts of a society.

6.3 Scope, context and criteria

6.3.1 General

The guidance provided in ISO 31000:2018, 6.3.1 applies.

In addition to the guidance provided in ISO 31000:2018, 6.3.1, for organizations using AI the scope of the AI risk management, the context of the AI risk management process and the criteria to evaluate the significance of risk to support decision-making processes should be extended to identify where AI

systems are being developed or used in the organization. Such an inventory of AI development and use should be documented and included in the organization's risk management process.

6.3.2 Defining the scope

The guidance provided in ISO 31000:2018, 6.3.2 applies.

The scope should take the specific tasks and responsibilities of the different levels of an organization into account. Moreover, the objectives and purpose of the AI systems developed or used by the organization should be considered.

6.3.3 External and internal context

The guidance provided in ISO 31000:2018, 6.3.3 applies.

Because of the magnitude of potential effects of AI systems, the organization should pay special attention to the environment of its stakeholders when forming and establishing the context of the risk management process.

Care should be taken to consider a list of stakeholders, including, but not limited to:

— the organization (itself);

— customers, partners and third parties;

— suppliers;

— end users;

— regulators;

— civil organizations;

— individuals;

— affected communities;

— societies.

Some other considerations for external and internal context are:

— whether AI systems can harm human beings, deny essential services (which if interrupted would endanger life, health or personal safety) or infringe human rights (e.g. by unfair and biased automated decision-making) or contribute to environmental harm;

— external and internal expectations for the organization's social responsibility;

— external and internal expectations for the organization's environmental responsibility.

The guidelines in ISO 26000:2010[2] outlining aspects of social responsibility should apply as a framework for understanding and treating risk, particularly on core subjects of organizational governance, human rights, labour practices, the environment, fair operating practices, consumer issues and community involvement and development.

NOTE Further background information on trustworthiness is available in ISO/IEC TR 24028:2020[3].

6.3.4 Defining risk criteria

The guidance provided in ISO 31000:2018, 6.3.4 applies.

In addition to the guidance provided in ISO 31000:2018, 6.3.4, Table 4 provides additional guidelines on factors to be considered when defining risk criteria:

Table 4 — Additional guidance when defining risk criteria

Considerations for defining risk criteria as provided in ISO 31000:2018, 6.3.4	Additional considerations in the context of the development and use of AI systems
— The nature and type of uncertainties that can affect outcomes and objectives (both tangible and intangible);	— Organizations should take reasonable steps to understand uncertainty in all parts of the AI system, including the utilized data, software, mathematical models, physical extension, and human-in-the-loop aspects of the system (such as any related human activity during data collection and labelling).
— How consequences (both positive and negative) and likelihood will be defined and measured;	
— Time-related factors;	— No specific guidance beyond ISO 31000:2018
— Consistency in the use of measurements;	— Organizations should be aware that AI is a fast-moving technology domain. Measurement methods should be consistently evaluated according to their effectiveness and appropriateness for the AI systems in use.
— How the level of risk is to be determined;	— Organizations should establish a consistent approach to determine the risk level. The approach should reflect the potential impact of AI systems regarding different AI-related objectives (see Annex A).
— How combinations and sequences of multiple risks will be taken into account;	— No specific guidance beyond ISO 31000:2018
— The organization's capacity.	— The organization's AI capacity, knowledge level and ability to mitigate realized AI risks should be considered when deciding its AI risk appetite.

6.4 Risk assessment

6.4.1 General

The guidance provided in ISO 31000:2018, 6.4.1 applies.

AI risks should be identified, quantified or qualitatively described and prioritized against risk criteria and objectives relevant to the organization. Annex B provides a sample catalogue of AI-related risk sources. Such a sample catalogue cannot be considered comprehensive. However, experience has shown the value of using such a catalogue as base for any organization performing a risk assessment exercise for the first time or integrating AI risk management into existing management structures. The catalogue serves as a documented baseline for these organizations.

Organizations engaged in the development, provisioning or application of AI systems therefore should align their risk assessment activities with the system life cycle. Different methods for risk assessment can apply to different stages of the system life cycle.

6.4.2 Risk identification

6.4.2.1 General

The guidance provided in ISO 31000:2018, 6.4.2 applies.

6.4.2.2 Identification of assets and their value

The organization should identify assets related to the design and use of AI that fall within the scope of the risk management process as defined in 6.3.2. Understanding what assets are within the scope and the relative criticality or value of those assets is integral to assessing the impact. Both the value of the asset and the nature of the asset (tangible or intangible) should be considered. Additionally, in relation

to the development and use of AI, assets should be considered in the context of elements including but not limited to the following:

— Assets of and their value to the organization:
  — Tangible assets can include data, models and the AI system itself.
  — Intangible assets can include reputation and trust.
— Assets of and their value to individuals:
  — Tangible assets can include an individual's personal data,
  — Intangible assets can include privacy, health, and safety of an individual.
— Assets of and their value to communities and societies:
  — Tangible assets can include the environment,
  — Intangible assets are likely more values based, such as socio-cultural beliefs, community knowledge, educational access and equity.

For valuation of assets and the relation to impact, see 6.4.2.6 and 6.4.3.2.

NOTE The use of the word "asset" with the illustrative examples in this clause does not have any legal implications.

6.4.2.3 Identification of risk sources

The organization should identify a list of risk sources related to the development or use of AI, or both, within the defined scope.

Risk sources can be identified within, but not limited to, the following areas:

— organization;
— processes and procedures;
— management routines;
— personnel;
— physical environment;
— data;
— AI system configuration;
— deployment environment;
— hardware, software, network resources and services;
— dependence on external parties.

Examples of AI-related risk sources can be found in Annex B.

6.4.2.4 Identification of potential events and outcomes

The organization should identify potential events that are related to the development or use of AI and can result in a variety of tangible or intangible consequences.

Events can be identified through one or more of the following methods and sources:

— published standards;

— published technical specifications;

— published technical reports;

— published scientific papers;

— market data on similar systems or applications already in use;

— reports of incidents on similar systems or applications already in use;

— field trials;

— usability studies;

— the results of appropriate investigations;

— stakeholder reports;

— interviews with, and reports from, internal or external experts;

— simulations.

6.4.2.5 Identification of controls

The organization should identify controls relevant to either the development or use of AI, or both. Controls should be identified during the risk management activities and documented (in internal systems, procedures, audit reports, etc.).

Controls can be utilized to positively affect the overall risk by mitigating risk sources and events and outcomes.

The operating effectiveness of the identified controls should also be taken into account, particularly control failures.

6.4.2.6 Identification of consequences

As part of AI risk assessment, the organization should identify risk sources, events or outcomes that can lead to risks. It should also identify any consequences to the organization itself, to individuals, communities, groups and societies. Organizations should take particular care to identify any differences between the groups who experience the benefits of the technology and the groups who experience negative consequences.

Consequences to the organization necessarily differ from those to individuals and to societies. Consequences to organizations can include but are not limited to:

— investigation and repair time;

— (work) time gained and lost;

— opportunities gained or lost;

— threats to health or safety of individuals;

— financial costs of specific skills to repair the damage;

— employee recruitment, satisfaction and retention;

— image reputation and goodwill;

— penalties and fines;

— customer litigations.

Depending on the context, consequences to individuals and to societies can be more general, in which case the organization can be unable to estimate exactly what the effect to each individual or to societies is.

Rather than specifying each type of effect, this can be considered generally as with the degree of the criticality of effects (for example, to privacy, fairness, human rights, etc., in the case of an individual, or to the environment in the case of societies) being a key element.

The exact effects can depend on the context in which the organization operates and areas for which the AI system is developed and used.

NOTE 1 Consequences can be positive or negative. The organization can consider both when assessing the consequences to the organization, to individuals and to societies.

NOTE 2 Consequences to individuals and societies usually can also lead to consequences to the organization. For example, a safety incident to a user of a product of the organization can result in liability claims to the organization and negatively impact its reputation and product sales.

6.4.3 Risk analysis

6.4.3.1 General

The guidance provided in ISO 31000:2018, 6.4.3 applies.

The analysis approach should be consistent with the risk criteria developed as part of establishing the context (see 6.3).

6.4.3.2 Assessment of consequences

When assessing the consequences identified in the risk assessment, the organization should distinguish between a business impact assessment, an impact assessment for individuals and a societal impact assessment.

Business impact analyses should determine the degree to which the organization is affected, and consider elements including but not limited to the following:

— criticality of the impact;

— tangible and intangible impacts;

— criteria used to establish the overall impact (as determined in 6.3.4).

Impact analyses for individuals should determine the degree to which an individual can be affected by the development or use of AI by the organization, or both. They should consider elements including but not limited to the following:

— types of data used from the individuals;

— intended impact of the development or use of AI;

— potential bias impact to an individual;

— potential impact on fundamental rights that can result in material and non-material damage to an individual;

— potential fairness impact to an individual;

— safety of an individual;

— protections and mitigating controls around unwanted bias and unfairness;

— jurisdictional and cultural environment of the individual (which can affect how relative impact is determined).

Impact analyses for societies should determine the degree to which societies can be affected by the either development or use of AI by the organization, or both. They should consider elements including but not limited to the following:

— scope of societal impact (how broad is the reach of the AI system into different populations), including who the system is being used by or designed for (for instance, governmental use can potentially impact societies more than private use);

— how an AI system affects social and cultural values held by various affected groups (including specific ways that the AI system amplifies or reduces pre-existing patterns of harm to different social groups).

6.4.3.3 Assessment of likelihood

Where applicable, the organization should assess the likelihood of occurrence of events and outcomes causing risks. Likelihood can be determined on a qualitative or quantitative scale and should align to the criteria established as part of 6.3.4. Likelihood can be informed and affected by (not limited to):

— types, significance, and number of risk sources;

— frequency, severity, and pervasiveness of threats;

— internal factors such as operational success of policies and procedures and motivation of internal actors;

— external factors such as geography and other social, economic and environmental concerns;

— success (mitigation) or failure of controls (see 6.4.2.5).

Organizations should incorporate likelihood calculations only where they are applicable and useful for identifying where to apply risk treatments. There can be significant technical, economic and heuristic issues with decision-making based likelihoods, particularly when the likelihood either can't be calculated or where the calculation has a large margin of error.

6.4.4 Risk evaluation

The guidance provided in ISO 31000:2018, 6.4.4 applies.

6.5 Risk treatment

6.5.1 General

The guidance provided in ISO 31000:2018, 6.5.1 applies.

6.5.2 Selection of risk treatment options

The guidance provided in ISO 31000:2018, 6.5.2 applies.

Risk treatment options defined by the organization should be designed to reduce negative consequences of risks to an acceptable level, and to increase the likelihood that positive outcomes can be achieved. If the required reduction of negative outcomes cannot be achieved by applying different risk treatment options, the organization should carry out a risk-benefit analysis for the residual risks.

In accordance with ISO 31000:2018, 6.5.2 the organization should consider:

— avoiding the risk by deciding not to start or continue with the activity that gives rise to the risk;

— taking or increasing the risk in order to pursue an opportunity;

— removing the risk source;

— changing the likelihood;

— changing the consequences;

— sharing the risk (for instance, through contracts or buying insurance);

— retaining the risk by informed decision.

6.5.3 Preparing and implementing risk treatment plans

The guidance provided in ISO 31000:2018, 6.5.3 applies.

Once the risk treatment plan has been documented, the risk treatment measures selected in 6.5.2 should be implemented.

The implementation of each risk treatment measure and its effectiveness should be verified and recorded according to 6.7.

6.6 Monitoring and review

The guidance provided in ISO 31000:2018, 6.6 applies.

6.7 Recording and reporting

The guidance provided in ISO 31000:2018, 6.7 applies.

The organization should establish, record, and maintain a system for the collection and verification of information on the product or similar products from the implementation and post-implementation phases. The organization should also collect and review publicly available information on similar systems on the market.

This information should then be assessed for possible relevance on the trustworthiness of the AI system. In particular, the evaluation should assess whether previously undetected risks exist or previously assessed risks are no longer acceptable. This information can be fed and factored into the organization's AI risk management process as adjustment of objectives, use cases or lessons learned.

If any of these conditions apply, organizations should perform the following:

— assess the effect on previous risk management activities and feed the results of this assessment back into the risk management process.

— carry out a review of the risk management activities for the AI system. If there is a possibility that the residual risk or their acceptance have changed, the effects on existing risk control measures should be evaluated.

The results of this assessment should be recorded. The risk management record should allow the traceability of each identified risk through all risk management processes. The records can leverage a common template that is agreed upon by the organization.

In addition to the documentation of the scope, context and criteria (see 6.3), risk assessment (see 6.4) and risk treatment (see 6.5), the record should include at least the following information:

— a description and identification of the system that has been analysed;

— methodology applied;

— a description of the intended use of the AI system;

— the identity of the person(s) and organization that carried out the risk assessment;

— the terms of reference and date of the risk assessment;

— the release status of the risk assessment;

— if and to what degree objectives have been met.

Annex A
(informative)

Objectives

A.1 General

When identifying risks of AI systems, various AI-related objectives should be taken into account, depending on the nature of the system under consideration and its application context. AI-related objectives to consider include but are not limited to the objectives described in Clauses A.2 to A.12.

A.2 Accountability

Accountability refers both to a characteristic of organizations and to a system property:

— Organizational accountability means that an organization takes responsibility for its decisions and actions by explaining them and being answerable for them to the governing body, to legal authorities and more broadly to stakeholders.

— System accountability relates to being able to trace the decisions and actions of an entity to that entity.

The use of AI can change existing accountability frameworks. Where previously persons performed actions for which they would be held accountable, such actions can now be fully or partially performed by AI systems. Who would be accountable in this case is an ongoing consideration by regulators. Developers and users of AI systems should be aware of the related legislation in the countries where the system is brought onto the market and used.

A.3 AI expertise

AI systems and their development are different from non-AI software solutions. A selection of dedicated specialists with inter-disciplinary skillsets and expertise in assessing, developing and deploying AI systems is needed. Organizations should ensure that people with such expertise are engaged in the development and specification of AI systems.

Expertise of AI should extend to the end users of AI systems. Users should have sufficient understanding of how the AI system functions and are empowered to detect and override erroneous decisions or outputs.

A.4 Availability and quality of training and test data

AI systems based on ML need training and test data in order to train and verify the systems for the intended behaviour. The deployed AI system operates on production data. The training, test and production data should be fit to the intended behaviour with respect to data type and quality.

Training and test data should be validated for their currency and relevance for the intended purpose. The amount of training and test data required can vary based on the intended functionality and complexity of the environment. The training and test data should have sufficiently diverse features in order to provide strong predictive power for the AI system. Furthermore, consistency should be ensured across training and test data, while using independent datasets when applicable.

It is possible that training and test data is not available in the company and is sourced externally. Data quality should be ensured also in that case.

A.5 Environmental impact

The use of AI can cause effects from an environmental point of view. The use of AI can have positive effects on the environment. For example, an AI system can be used to reduce nitrogen oxide in a gas turbine. The use of AI can also have a negative effect on the environment due to the extensive use of resources. For example, the training phase of some AI systems requires computing resources and can consume substantial amounts of electrical power. These impacts on the environment should be considered.

A.6 Fairness

The use of AI systems for automated decision-making can be unfair to specific persons or groups of persons. Unfair outcomes have a number of causes such as bias in objective functions, imbalanced data sets, and human biases in training data and in providing feedback to systems. Unfairness can also be caused by a bias issue in the product concept, the problem formulation or choices about when and where to deploy AI systems.

For further information on bias and fairness in AI systems, see ISO/IEC TR 24027[4].

A.7 Maintainability

Maintainability is related to the ability of the organization to handle modifications of the AI system in order to correct defects or adjust to new requirements. Because AI systems based on ML are trained and do not follow a rule-based approach, the maintainability of an AI system and its implications should be investigated.

A.8 Privacy

Privacy is related to the ability of individuals to control or influence what information related to them can be collected, stored and processed, and by whom that information can be disclosed. As explained in ISO/IEC TR 24028:2020[3], "many AI techniques (e.g. deep learning) highly depend on big data since their accuracy relies on the amount of data they use. The misuse or disclosure of some data, particularly personal and sensitive data (e.g. health records) can have harmful effects on data subjects. Thus, privacy protection has become a major concern in big data analysis and AI."

Consideration should be taken to determine if an AI system can infer sensitive personal data. For AI systems, protecting privacy includes protecting the data used for building and operating the AI system, ensuring that the AI system cannot be used to give unwarranted access to its data, and protecting access to models personalized for an individual or that can be used to infer information or characteristics of similar individuals.

Improper collections, uses and disclosures of personal information can also have direct impacts on fundamental human rights such as discrimination and freedom of expression and information. Impacts on ethic principles in terms of respect of human values, and human dignity should also be considered.

NOTE A data protection impact assessment (see ISO/IEC 29134:2017[5]), often referred to as a privacy impact assessment, is a useful tool for managing the risks related to the use of personal data during the collection of data, training of an AI system, and use of an AI system.

A.9 Robustness

Robustness is related to the ability of a system to maintain its level of performance under the various circumstances of its usage. The degree to which an AI system or related component can function correctly in the presence of invalid inputs or stressful environmental conditions should be taken into consideration as well as the ability to reproduce measures and results.

Robustness poses new challenges in the context of AI systems. Neural network architectures represent a specific challenge as they are both hard to explain and sometimes have unexpected behaviour due to their nonlinear nature. Characterizing the robustness of neural networks is an open area of research, and there are limitations to both testing and verification approaches.

For further information on robustness of neural networks, see ISO/IEC TR 24029-1[6].

A.10 Safety

The use of AI systems can introduce new safety threats. Safety relates to the expectation that a system does not, under defined conditions, lead to a state in which human life, health, property or the environment is endangered. Use of AI systems in automated vehicles, manufacturing devices, and robots can introduce risks related to safety. Specific standards for particular application domains (e.g. the design of machinery, transport, medical devices) should be taken into account for AI systems in these domains.

For further information on functional safety in AI systems, see ISO/IEC TR 5469[1] [7].

A.11 Security

Information security risk management is defined in ISO/IEC 27005:2022[8]. In the context of AI, and in particular with regard to AI systems based on ML approaches, several new issues such as data poisoning, adversarial attacks and model stealing as described in ISO/IEC TR 24028:2020[3] should be considered beyond classical information and system security concerns.

A.12 Transparency and explainability

Transparency relates both to characteristics of an organization operating AI systems, and to those systems themselves. Organizations are sometimes transparent on how they apply such systems, how they use collected data (such as consumer and user data, public data, other collected data sets), which measures they put in place to manage AI systems, understand and control their risks, etc. Transparency of AI systems is to provide appropriate information about a system (e.g. capabilities and limitations) to stakeholders to enable them to assess development, operation and use of AI systems against their objectives. AI system explainability relates to an ability to rationalize and help to understand how, for a specific system, its outcome has been generated.

1) Under preparation. Stage at the time of publication: ISO/IEC DTR 5469:2022.

Annex B
(informative)

Risk sources

B.1 General

When identifying risks of AI systems, various risks sources should be taken into account depending on the nature of the system under consideration and its application context. Risk sources to consider include but are not limited to the issues and opportunities described in Clauses B.2 to B.8.

B.2 Complexity of environment

The complexity of the environment[9] of an AI system determines the range of potential situations an AI system is intended to support in its operational context.

Certain AI technologies like ML are specifically suited to handle complex environments and are therefore often used for systems used for complex environments like automated driving. A great challenge however is to identify in the design and development process all relevant situations that the system is expected to handle and that the training and test data cover all these situations.

Hence, complex environments can result in additional risks relative to simple environments. Special consideration should be given to determining the degree to which the AI system environment is understood:

— Complete understanding of environment that is only possible for simple predictable or controlled environments, such that the AI system is prepared for all possible states of the environment that it can encounter, allows for better risk control.

— In case of partial understanding due to high complexity or uncertainty of the environment, such that the AI system cannot forecast all possible states of the environment (for instance, autonomous driving), it cannot be assumed that all relevant situations are considered. This can result in a level of uncertainty, which is a source of risk, and should be taken into account when designing such systems.

B.3 Lack of transparency and explainability

Transparency is about communicating appropriate activities and decisions of an organization (e.g. policies, processes) and appropriate information about an AI system (e.g. capabilities, performance, limitations, design choices, algorithms, training and test data, verification and validation processes and results) to relevant stakeholders. This can enable stakeholders to assess development, operation and use of AI systems against their expectations. The kind and level of information that is appropriate strongly depends on the stakeholders, use case, system type and legislative requirements. If organizations are unable to provide the appropriate information to the relevant stakeholders, it can negatively affect trustworthiness and accountability of the organization and AI system.

Explainability is the property of an AI system that the important factors influencing a decision can be expressed in a way that humans can understand. An ML model can have behaviour that is difficult to understand by inspection of the model or the algorithm used to train it, especially in the case of deep learning. If such important factors cannot be expressed, validation of the AI system and the trust of humans in the system are negatively affected as it is not clear why the system has made a decision and if it can make the correct decision in all cases. This uncertainty can result in many risks and strongly effect general objectives such as trustworthiness and accountability, and specific objectives such as