भारतीय मानक **IS/ISO/IEC 8183 : 2023**

Indian Standard

सूचना प्रौद्योगिकी — आर्टिफिशियल इंटेलिजेंस — डेटा लाइफ साईकिल फ्रेमवर्क

Information Technology — Artificial Intelligence — Data Life Cycle Framework

ICS 35.020

भारतीय मानक ब्यूरो
BUREAU OF INDIAN STANDARDS
मानक भवन, 9 बहादुर शाह ज़फर मार्ग, नई दिल्ली - 110002
MANAK BHAVAN, 9 BAHADUR SHAH ZAFAR MARG
NEW DELHI - 110002
www.bis.gov.in www.standardsbis.in

January 2024 **Price Group 7**

Artificial Intelligence Sectional Committee, LITD 30

NATIONAL FOREWORD

This Indian Standard which is identical to ISO/IEC 8183 : 2023 'Information technology — Artificial intelligence — Data life cycle framework' issued by the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) jointly was adopted by the Bureau of Indian Standards on the recommendations of the Artificial Intelligence Sectional Committee and approval of the Electronics and Information Technology Division Council.

The text of ISO/IEC standard has been approved as suitable for publication as an Indian Standard without deviations. Certain conventions are however not identical to those used in Indian Standards. Attention is particularly drawn to the following:

a) Wherever the words 'International Standard' appears referring to this standard, they should be read as 'Indian Standard': and

b) Comma (,) has been used as a decimal marker while in Indian Standards, the current Practice is to use a point (.) as the decimal marker.

In this adopted standard, reference appears to certain International Standards for which Indian Standards also exist. The corresponding Indian Standards, which are to be substituted in their places, are listed below along with their degree of equivalence for editions indicated. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document applies, including any corrigenda and amendment:

International Standards	*Corresponding Indian Standard*	*Degree of Equivalence*
ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	IS/ISO/IEC 22989 : 2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	Identical

Contents

Page

Introduction ..... **v**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**

4 Symbols and abbreviated terms **1**

5 Data life cycle overview **1**

6 Data life cycle framework **2**

6.1 General 2
6.2 Stage 1: Idea conception 3
6.3 Stage 2: Business requirements 4
6.4 Stage 3: Data planning 4
6.5 Stage 4: Data acquisition 5
6.6 Stage 5: Data preparation 5
6.7 Stage 6: Building a model 6
6.8 Stage 7: System deployment 6
6.9 Stage 8: System operation 7
6.10 Stage 9: Data decommissioning 7
6.11 Stage 10: System decommissioning 7

7 Stages and processes within the data life cycle **7**

Bibliography **10**

Introduction

Artificial intelligence (AI) systems are being adopted by organizations of all types, sizes and purposes. Data are essential to the development and operation of AI systems.

In the field of AI systems, there are many data life cycles in use and under consideration for different purposes (e.g. data quality, bias in data, data governance, development and use of AI systems). Without an overarching framework, these different data life cycles can be challenging to correctly interpret by those without previous knowledge, context and expertise. There is a risk that these multiple data life cycles will not be applied as intended.

This document provides a data life cycle overview in Clause 5, describes a data life cycle framework in Clause 6 and provides more information on the stages or processes of the data life cycle in Clause 7.

Indain Standard

INFORMATION TECHNOLOGY ARTIFICIAL INTELLIGENCE — DATA LIFE CYCLE FRAMEWORK

1 Scope

This document defines the stages and identifies associated actions for data processing throughout the artificial intelligence (AI) system life cycle, including acquisition, creation, development, deployment, maintenance and decommissioning. This document does not define specific services, platforms or tools. This document is applicable to all organizations, regardless of type, size or nature, that use data in the development and use of 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, *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/IEC 22989 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 Symbols and abbreviated terms

AI	artificial intelligence
DPIA	data protection impact assessment
JSON	JavaScript object notation
ML	machine learning
OWL	web ontology language
PII	personally identifiable information
XML	extensible markup language

5 Data life cycle overview

The data life cycle for AI systems encompasses the processing of data from the earliest conception of a new AI system to the eventual decommissioning of the system and is separated into a number of distinct stages. Each stage will often, but not always, be part of a data life cycle for an AI system.

A data life cycle represents all the stages through which data can pass within any system that uses data of any kind. It is designed to support the achievement of objectives related to system governance, system utility, data quality and data security, by ensuring that data processing is given due consideration during the planning, development, use and decommissioning of the system.

The detailed purpose and timing of use of these stages throughout the life cycle are influenced by multiple factors, including societal, commercial, organizational and technical considerations, each of which can vary or at times be combined with other stages during the life of a system. This document describes the following 10 stages:

— stage 1 – idea conception;

— stage 2 – business requirements;

— stage 3 – data planning;

— stage 4 – data acquisition;

— stage 5 – data preparation;

— stage 6 – building model;

— stage 7 – system deployment;

— stage 8 – system operation;

— stage 9 – data decommissioning;

— stage 10 – system decommissioning.

For information about a data life cycle for data usage, see ISO/IEC 5212: —[1)].

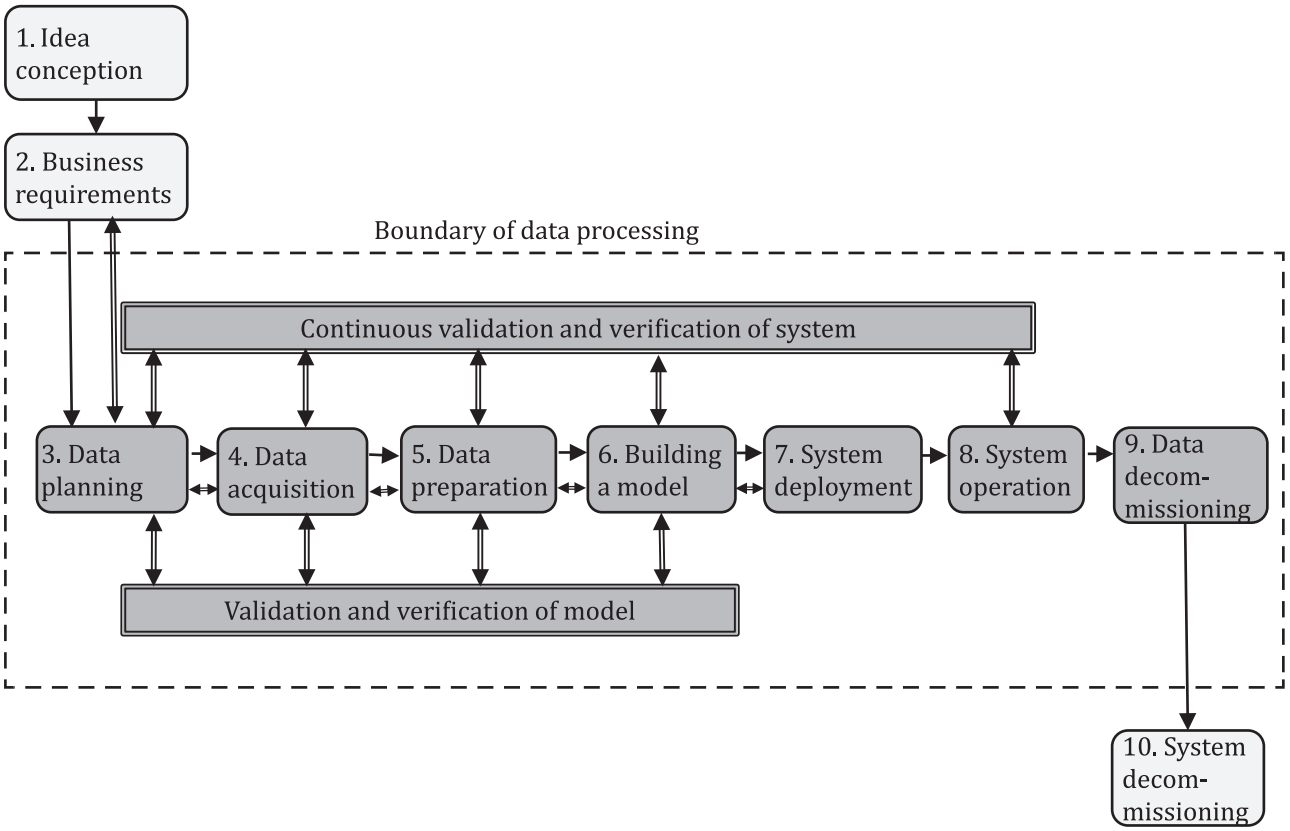
6 Data life cycle framework

6.1 General

The data life cycle framework, shown in Figure 1, identifies a set of conceptually distinct stages that data used in an AI system go through from data planning to data decommissioning. Figure 1 also includes idea conception, business requirements and system decommissioning, which are system-level life cycle stages. For information regarding data sets, refer to ISO/IEC 23053:2022, 6.5. Life cycle processes appropriate to a defined task can be assigned to each stage. Life cycle processes describe the actions taken on the data within the life cycle stage.

Stage 9 (data decommissioning) and stage 10 (system decommissioning) both pertain to decommissioning but stage 9 specifically covers what happens to the data (e.g. secure deletion, archiving, repurposing) while stage 10 covers what happens to the system irrespective of what happens to the data that is being processed.

1) Under preparation. Stage at the time of publication: ISO/IEC DIS 5212:2023.



Key

primary development pathway

feedback pathway

stage with data processing

stage outside the data processing boundary

NOTE 1 The single-headed arrows depict a linear path through the life cycle stages, while the double-headed arrows show feedback paths between life cycle stages.

NOTE 2 The verification and validation of the model refers to the internal development process, whose output is a model. The validation and verification of the system refers to the system as a whole, extending through its entire period of operation.

Figure 1 — Data life cycle framework

6.2 Stage 1: Idea conception

Idea conception is when a need or requirement for a new or revised AI system is recognized. The AI system can be used as a partial or complete solution to an existing or potential problem or opportunity faced by the organization.

Idea conception can also be driven by broader organizational context needs (e.g. economic, technical, strategic, market or legal requirements). Ultimately, this idea can be expressed as one or more questions that the AI system can answer. The questions to which the AI system provides answers should be mapped to and aligned with business objectives and metrics.

6.3 Stage 2: Business requirements

The business requirements stage can include one or more stakeholders with appropriate authority or influence deciding to investigate whether the idea can be turned into a functioning system and deciding whether to invest further in the idea. This stage involves:

— determining the ambition of the project (e.g. vision, goals and strategy);

— determining assets, including those available and those that need to be acquired;

— specifying the data requirements, a key element for AI systems, based on the business goals and end user requirements;

— identifying enablers for the project, including in-house skills and knowledge, organizational architecture, technology and external resources;

— ensuring the project can be developed in line with organizational policies and procedures (or processes), including:

  — compliance (e.g. privacy requirements);

  — ethics (e.g. fairness of outcomes);

  — culture;

  — leadership;

  — governance processes.

The business requirements stage can conclude with a determination of whether or not the project is feasible.

NOTE The business requirements stage does not include processing of the data.

6.4 Stage 3: Data planning

The data planning stage involves deciding upon the scope of the data required to address the questions identified in the business requirements stage. The primary data factors for consideration at this stage include:

— whether the necessary data exist, are available for reuse, need to be acquired, collected, transformed, authored or curated, or a combination of some or all of these;

— amount of data required;

— source of the data;

— whether synthetic, i.e. artificial, data can be created to augment available data;

— what data outputs are created and how the system will deal with them;

— format of the data;

— what the data represent;

— properties of the data that can affect the choice of algorithm for model development;

— data licensing requirements;

— data security, privacy and resilience requirements;

— data acquisition requirements, such as data collection requirements;

— data safety requirements;

— data storage and retention requirements (e.g. type, cost, capacity, performance, embedded capabilities, timelines for deletion).

The aim of this stage is to ensure that the goals, requirements and needs established in the idea conception and business requirements stages can be met. Candidate data sets can be identified, acquired and examined. Data sets can be internal or sourced from public sources, government bodies, vendors or third-party providers.

6.5 Stage 4: Data acquisition

The data acquisition stage involves the creation of, or access to the data identified in the data planning stage. Acquired data can come from internal sources, third parties or the community (e.g. open data, public domain data).

Acquiring data from third parties can involve consent, contracts and licences, as described in 6.4.

Data can be in different forms [e.g. static, streaming content, real-time internet of things (IoT) data] and different formats (e.g. XML, JSON, delimited text, binary). Data can be structured, semi-structured or unstructured.

Data acquisition processes should make use of good management practices (e.g. security, privacy, quality).

6.6 Stage 5: Data preparation

The data preparation stage involves the processing of the data gathered in the data acquisition stage. The data preparation stage can involve the use of the following operations:

— Decryption: transforming encrypted data to a state where it can be used in an AI system, when needed and possible.

— Cleaning: includes transforms and operations such as relevance validation, deduplication, outlier removal, bias mitigation, imputation of missing entries, correcting entries and correcting data formats.

— Feature engineering: appropriate features can be selected for use and possible transformation to improve the performance of machine learning. Existing features can be combined and processed to produce new features that can improve training and inference.

— Normalizing and scaling: it can be necessary to transform data of widely differing ranges to fit within a specified range [e.g. between 0 and 1 (unit norm)]. It can also be necessary to scale the data samples to fit in a standard distribution or other specified distribution.

— Data organization: data can be reorganized without changing its value or meaning. It can be necessary to append data sets or join tables to obtain a complete data set. It can be necessary to combine or split columns or fields to achieve a specified data set structure.

— Labelling: the values of target variables should be established by a suitable manual or automatic process. For example, with supervised learning, values can be determined manually. With semi-supervised learning, values can be determined through automated techniques. An example of manual data labelling for an image recognition application is the use of humans to determine the species of animals in a set of digital images.

— Enrichment: running tools to link diverse data sources and to add additional context to the data. For example, unstructured data can be processed by natural language processing (NLP) tools to extract named entities. Place names and addresses can be identified and geocoded using a gazetteer to enable location-based analysis later.

— De-identification: it can be necessary to remove personally identifiable information from the data to protect the privacy of data subjects.

— Resampling: for example, it can be useful to sub-sample large data sets to improve the consistency of statistical significance of different data classes or to reduce the time required to build and test a model while achieving useful results. Similarly, data sets can be up-sampled (i.e. sampled with replacement) to improve the consistency of statistical significance of different data classes.

— Encoding: it can be necessary to encode data used to build a model. For example, it can be necessary to encode text values for categorical variables (e.g. converting text features to numerical features, digitizing analogue signals).

— Integrity verification: applying a process specific to the kind of data to check the overall integrity of a data set. This is more likely to be applicable to structured and semi-structured data, for which there can already be a structural model (e.g. database schema, formal ontology).

— Data provenance: updating the provenance record of each data set to record changes and operations undertaken.

— Data anonymization or pseudonymization.

NOTE For additional information on data preparation for ML, see ISO/IEC 23053.

6.7 Stage 6: Building a model

Building a model involves deciding upon the organization, storage and access to the data so that it can be processed to build a model that delivers some function in support of the business goals. The model-building process can be a distinct activity with a conclusion, whose output is a fixed model or a continuous activity, whereby the model is subject to ongoing revision (continuous learning). In either case, the model can be either of the following:

a) The result of training an ML algorithm using training data. Model building can take place centrally or across a network of resources (e.g. federated learning, split learning). Examples of trained models include decision trees, inductive logic programming and various types of neural networks.

b) The result of combining human engineered knowledge (e.g. declarative or procedural) with an inference process. Examples of forms of human engineered knowledge include Horn clauses (e.g. as used in the programming language Prolog), varieties of description logic (e.g. as used in OWL and OWL2) and answer set programming.

NOTE 1 Inductive logic programming is a form of machine learning over symbolic structures in which a logic program is modified automatically to satisfy specified goal conditions.

NOTE 2 Answer set programming is a form of logic programming that uses an answer set solving algorithm to construct a symbolic model in which all variables are replaced with literals.

Data are used to train and calibrate the model alongside human expertise and to verify that the system's outputs and performance meet stakeholder expectations. A data protection impact assessment (DPIA) can be conducted to deal with concerns such as privacy issues in the outputs. The model can also be evaluated for other potential issues (e.g. bias, fairness, other ethical issues) and remedial actions taken.

During the model-building stage, the model should be evaluated to ensure that it meets the requirements established in prior stages (e.g. business requirements, data planning, data acquisition). Likewise, the outputs and performance of the model should be evaluated against the expectations of appropriate stakeholders, including the stakeholder's ability to use the model in practice. In some instances, an independent safety assessment (ISA) can be required for public-safety-related systems.

6.8 Stage 7: System deployment

System deployment involves the AI system going live in a target environment. This stage is not necessarily a simple switching on of the system; instead, it can include a number of processes testing, improving or ensuring that the system is operating as expected. As part of this stage, data flows should be examined to ensure they are working as envisaged, especially if the target environment included new systems or connections.

6.9 Stage 8: System operation

System operation involves the model generating outputs from input or production data. Output data can be processed in a number of ways, including some or all of the following:

— data ingestion or retrieval;

— data pseudonymization or anonymization;

— data manipulation or combination;

— data analysis;

— data visualization;

— data transmission;

— data storage.

This stage can include data access authorization, authentication and intended use.

Production data should be continuously monitored to ensure data quality is maintained and the system is not used for other than its original intended purpose.

Continuous verification and validation can be useful in reducing system operation risks but are not always feasible or appropriate. When feasible and appropriate, continuous verification and validation of the system should take place to ensure business requirements and appropriate stakeholder expectations are met. The system can be continuously improved as needed. New training data can be used to mitigate performance degradation. New training data should be managed per stage 3 (data acquisition) and stage 4 (data preparation).

6.10 Stage 9: Data decommissioning

Data decommissioning involves the disposition of data no longer used by the system (e.g. secure deletion, archiving, repurposing). Data categories should be defined and some categories of data should be retained for auditing purposes (e.g. log data to prove compliance)

The model itself can be subject to decommissioning if it contains residual elements of the training data or due to other requirements (e.g. security, privacy and confidentiality). Data licensing requirements can require deletion of data to a specified level. Additionally, deletion of the data containing PII can be required, for example, by applicable legal or contractual requirements.

6.11 Stage 10: System decommissioning

System decommissioning includes the cessation of data processing and the disposition of system components not covered by data decommissioning, such as the target environment components. Data not specific to the model, such as system logs, can be retained for future study or securely deleted.

7 Stages and processes within the data life cycle

Figure 2 shows the processes that are important to the data life cycle.

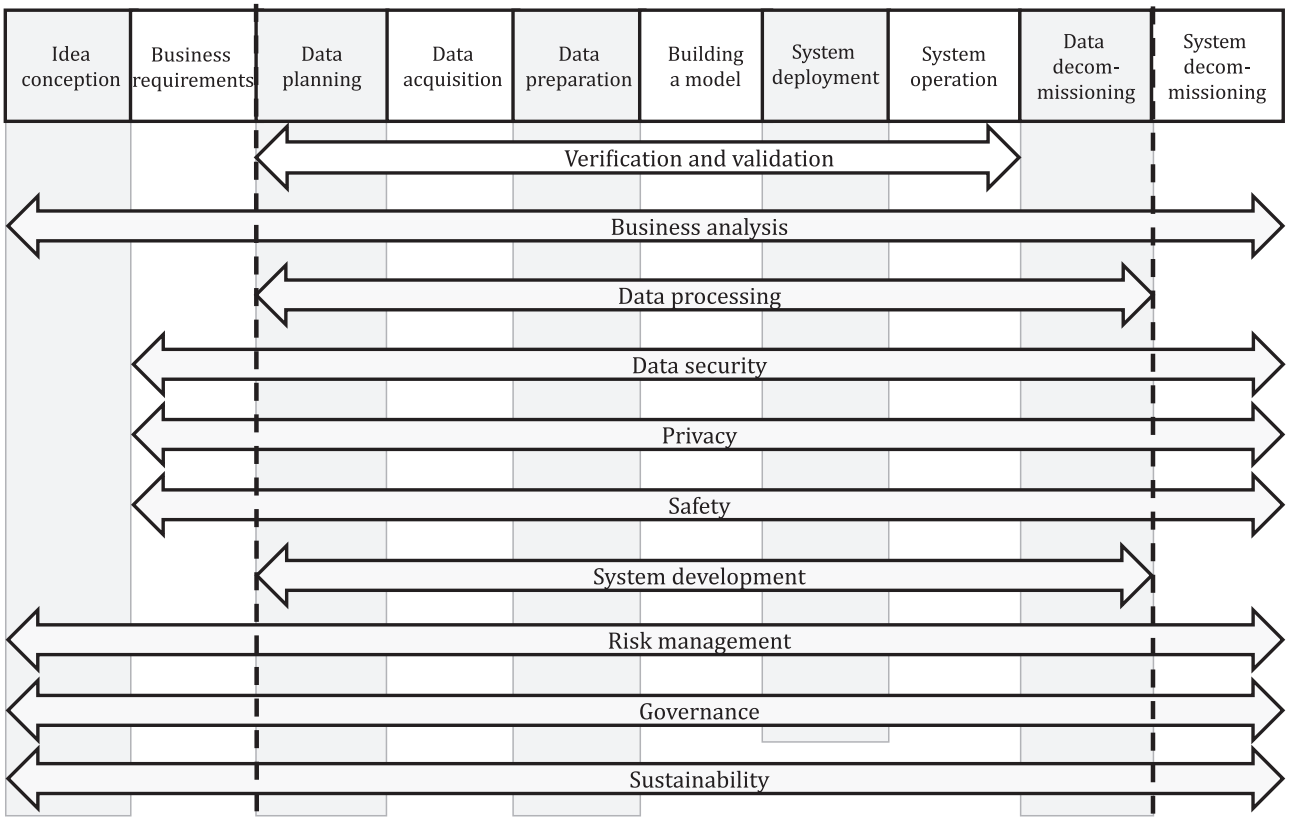


Figure 2 — A visual demonstration of how the processes relate to each stage

Figure 2 shows the processes and cross-cutting aspects that are important to the data life cycle. The processes include the following:

— Validation and verification: the two questions “Have we built the right system?” and “Have we built the system right?” are tested constantly by examining the outputs of the system against business requirements.

— Business analysis: the relationship between requirements of the organization and the ability of the AI system, and specifically the data used in the AI system, to deliver against those requirements is analysed at various points throughout the data life cycle.

— Data processing: every time data are touched they are processed; when they are collected, manipulated, stored, shared or deleted they are being processed. This has added significance for personal data and especially for sensitive categories of personal data. The provenance records of the data used in the AI systems should be updated as required by the organization.

— Data security: the data in an AI system should be kept confidential (i.e. secure from unauthorized access), accessible (i.e. available for authorized access) and with integrity (i.e. safe from unauthorized alteration).

— Privacy: PII should be protected and its integrity and availability maintained by applying the confidential, integrity and availability principle. Processes should incorporate PII protection, as well as mitigating accidental creation of new PII.

— Safety: ensuring the efficiency and effectiveness of the handling of data by the system to mitigate potential risks or harm that can be introduced by the AI system.

EXAMPLE 1 Data labelling: an AI system that is trained using a set of example data to age-label visual content correctly. In production the AI system age-labels an element of visual content that it was not trained to identify incorrectly, putting the future audience at risk.

EXAMPLE 2 Data monitoring: continuous monitoring of the data which are input to analysis is necessary to capture changes to data that the analysis cannot handle. If the initial conditions are wrong then possible outcomes are non-termination or meaningless results. Analysis can fail because the system is presented with data that do not satisfy properties that it is supposed to handle.

EXAMPLE 3 Legal requirements or assurance: where the data can be used in an AI application covered by specific legal or public safety requirements (e.g. rail, aviation, nuclear, medical, or oil and gas) can cause additional data assurance requirements.

NOTE Data assurance is agreed with or assessed by a regulator. This can be done within functional safety assessment or a broader assessment. Assurance can be performed for both static data and ongoing and dynamic data aspects.

— System development: creating a system that can be used to deliver against the organization's business requirements, including continuous validation and verification.

— Risk management: identification and management of any risks associated with the system, based on the risk appetite of the organization. Risk to the privacy of personal data and the security of the system and the data within it are included here.

— Governance: the role of the governing body and managers in operating and decommissioning the AI system and the data used within it. This includes a consideration of ethics, compliance with legal requirements, standards and best practice, accountability, risk management and fiduciary duty (see ISO/IEC 38507).

— Sustainability: the social and environmental implications of developing and using the AI system are considered so that the data processed in the system can be processed in as sustainable a fashion as possible. This is relevant both to the physical operation of the system (e.g. ensuring that operation is possible at periods of lower energy demand or that the system is protected against unnecessary use) and to the outputs of the system (e.g. that ethical standards are maintained).

Bibliography

[1] ISO/IEC 5212:—[2)], *Information technology — Data usage — Guidance for data usage*

[2] ISO/IEC 23053, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

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

2) Under preparation. Stage at the time of publication: ISO/IEC DIS 5212:2023.

Bureau of Indian Standards

BIS is a statutory institution established under the *Bureau of Indian Standards Act*, 2016 to promote harmonious development of the activities of standardization, marking and quality certification of goods and attending to connected matters in the country.

Copyright

Review of Indian Standards

Amendments are issued to standards as the need arises on the basis of comments. Standards are also reviewed periodically; a standard along with amendments is reaffirmed when such review indicates that no changes are needed; if the review indicates that changes are needed, it is taken up for revision. Users of Indian Standards should ascertain that they are in possession of the latest amendments or edition by referring to the website-www.bis.gov.in or www.standardsbis.in.

This Indian Standard has been developed from Doc No.: LITD 30 (23529).

Amendments Issued Since Publication

Amend No.	Date of Issue	Text Affected

BUREAU OF INDIAN STANDARDS

Headquarters:

Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002
Telephones: 2323 0131, 2323 3375, 2323 9402 *Website*: www.bis.gov.in

Regional Offices:		*Telephones*
Central	: 601/A, Konnectus Tower -1, 6th Floor, DMRC Building, Bhavbhuti Marg, New Delhi 110002	2323 7617
Eastern	: 8th Floor, Plot No 7/7 & 7/8, CP Block, Sector V, Salt Lake, Kolkata, West Bengal 700091	2367 0012 2320 9474
Northern	: Plot No. 4-A, Sector 27-B, Madhya Marg, Chandigarh 160019	265 9930
Southern	: C.I.T. Campus, IV Cross Road, Taramani, Chennai 600113	2254 1442 2254 1216
Western	: Plot No. E-9, Road No.-8, MIDC, Andheri (East), Mumbai 400093	2821 8093

Branches : AHMEDABAD. BENGALURU. BHOPAL. BHUBANESHWAR. CHANDIGARH. CHENNAI. COIMBATORE. DEHRADUN. DELHI. FARIDABAD. GHAZIABAD. GUWAHATI. HIMACHAL PRADESH. HUBLI. HYDERABAD. JAIPUR. JAMMU & KASHMIR. JAMSHEDPUR. KOCHI. KOLKATA. LUCKNOW. MADURAI. MUMBAI. NAGPUR. NOIDA. PANIPAT. PATNA. PUNE. RAIPUR. RAJKOT. SURAT. VISAKHAPATNAM.

Published by BIS, New Delhi