सत्यमेव जयते
भारत सरकार
Government of India

संचार मंत्रालय	Ministry of Communications
दूरसंचार विभाग	Department of Telecommunications
दूरसंचार अभियांत्रिकी केंद्र	Telecommunications Engineering Centre
अभिषरण एवं प्रसारण प्रभाग	Convergence & Broadcasting Division

File No: 14-1/2023-C&B/TEC/AI-Robustness/3416512 Dated 06-06-2025

Subject: Stakeholders consultation on "Draft Standard for Robustness Assessment and Rating of Artificial Intelligence Systems in Telecom Networks and Digital Infrastructure"

The formulation of "Standard for Robustness Assessment and Rating of Artificial Intelligence Systems in Telecom Networks and Digital Infrastructure" is being taken up.

2. Therefore in exercise of the powers conferred by rule 5(1) of the Telecommunications (Framework to Notify Standards, Conformity Assessment and Certification) Rules 2025, a "Draft Standard for Robustness Assessment and Rating of Artificial Intelligence Systems in Telecom Networks and Digital Infrastructure" is enclosed herewith (**Annexure-l**) for stakeholder consultation. It is requested to go through the aforesaid enclosed draft standard and offer your inputs/comments. The comments may please be furnished in the template sheet enclosed herewith as **Annexure-II**.

3. Comments may be sent by Email with subject "Comment on Draft Standard for Robustness Assessment and Rating of Artificial Intelligence Systems in Telecom Networks and Digital Infrastructure" to **jto-cb@gov.in,** with copies to **dircb2.tec-dot@gov.in.** and **ddgcb.tec@gov.in**

Enclosures: As above (i) Annexure-l and (ii) Annexure-II

Digitally signed by
Ram Raj Yadava
Date: 06-06-2025
15:59:32

R.R.Yadava
Director (Convergence & Broadcasting)

To,
All stakeholders,

Copy to:1. Sr DDG TEC - for kind information
2. AD(IT), TEC - with request for uploading on TEC website/Portal
3. AD(IMP&TEP), TEC - with request for uploading on TBT portal

Annexure-I

टीईसी का प्रारूप मानक दस्तावेज सं: टीईसी **57070:2025**

STANDARD DOCUMENT OF TEC
No. TEC DRAFT TEC 57070:2025

Robustness Assessment and Rating of Artificial Intelligence Systems in Telecom Networks and Digital Infrastructure

ISO 9001:2015

दूरसंचार अभियांत्रिकी केंद्र
खुर्शीदलाल भवन, जनपथ, नई दिल्ली -११०००१, भारत
TELECOMMUNICATION ENGINEERING CENTRE
KHURSHIDLAL BHAWAN, JANPATH, NEW DELHI–110001, INDIA
www.tec.gov.in

Release____: Month, Year

FOREWORD

Telecommunication Engineering Centre (TEC) is the technical arm of Department of Telecommunications (DOT), Government of India. Its activities include:

- Framing of TEC Standards for Generic Requirements for a Product/Equipment, Standards for Interface Requirements for a Product/Equipment, Standards for Service Requirements & Standard document of TEC for Telecom Products and Services
- Formulation of Essential Requirements (ERs) under Mandatory Testing and Certification of Telecom Equipment (MTCTE)
- Field evaluation of Telecom Products and Systems
- Designation of Conformity Assessment Bodies (CABs)/Testing facilities
- Testing & Certification of Telecom products
- Adoption of Standards
- Support to DoT on technical/technology issues

For the purpose of testing, four Regional Telecom Engineering Centers (RTECs) have been established which are located at New Delhi, Bangalore, Mumbai, and Kolkata.

ABSTRACT

This Standard enumerates detailed procedures for accessing and rating artificial intelligence systems for robustness. This standard outlines a comprehensive approach to evaluating the robustness of AI models in critical applications, focusing on metrics such as resilience to data shifts, integrity, reliability, explainability, transparency, privacy, and security. It provides a structured assessment methodology to identify vulnerabilities and propose mitigation strategies, ensuring that AI systems can withstand adversarial conditions and maintain consistent performance. Additionally, a rating methodology is introduced to quantify and benchmark the robustness of AI systems, offering telecom operators, developers, and policymakers a standardized approach to enhance trust and safety in AI-driven digital infrastructure.

A. HISTORY SHEET

S. No.	Standard No.	Equipment/Interface	Remarks
1.	TEC	Robustness Assessment and Rating of Artificial Intelligence Systems	

Contents

1.0 Introduction

The increasing integration of Artificial Intelligence (AI) and Machine Learning (ML) across various domains, including telecom networks and digital infrastructure, makes it essential to ensure their robustness. AI is now pivotal in optimizing network performance, automating operations, enhancing security, and improving user experiences. However, vulnerabilities in AI applications can lead to ethical, social, and legal issues. The National Digital Communications Policy-2018 emphasizes leveraging AI to enhance network quality, management, security, and reliability while mandating a holistic and harmonized approach to harnessing emerging technologies, including AI, through frameworks for testing and certification of new products and services. In this context, ensuring the robustness of AI systems is crucial for delivering consistent, reliable, and safe performance in real-world applications.

The concept of robustness in AI pertains to the system's ability to maintain stable and accurate performance despite variations in input data, potential adversarial attacks, and unforeseen operational conditions. This standard provides a comprehensive framework for assessing and enhancing the robustness of AI systems, specifically tailored for telecom and digital infrastructure contexts. Given the high-stakes nature of AI applications in these sectors, where even minor failures can lead to significant service disruptions, a structured approach to robustness assessment is necessary.

The standard provides a comprehensive view of robustness in the context of AI systems, which can be assured when the entire ecosystem (process and component) demonstrates six **Core Principles** and six **Core Elements** of AI robustness, identified in this standard, in its overall deployment and applications. It further provides key metrics associated with robustness, an assessment framework for evaluating these aspects, and a mitigation strategy to address identified risks and vulnerabilities. A rating methodology is proposed to provide a qualitative measure of an AI system's robustness, offering a reference scale for comparison. This framework aims to help stakeholders, including telecom operators, developers, and policymakers, evaluate, improve, and certify the robustness of AI systems deployed in critical infrastructure.

As organizations increasingly rely on AI-driven solutions for managing network operations and digital services, the demand for standardized procedures to assess and ensure the robustness of these systems becomes evident. The proposed standard can be applied through both **self-certification**, where internal assessments are conducted, and **independent certification** by external auditors. By establishing a clear and transparent process for robustness evaluation, the standard seeks to promote trust, reliability, and resilience in AI systems, ensuring their safe and effective deployment in the telecom sector.

While the standard focuses on telecom and digital infrastructure, it is important to note that the **core principles and core elements of robustness in AI are universally applicable**. This makes the standard relevant for other sectors and applications, where AI plays a critical role in optimizing operations, ensuring security, and delivering reliable performance. By addressing potential risks and vulnerabilities, this standard provides a robust and adaptable framework for promoting responsible AI practices across industries.

2.0 Scope, Limitations, and Users of the Standard

2.1. Scope & Limitations

This version of the standard encompasses the following areas within its scope:

1. Dimensions of robustness: This version of the standard covers availability, security, safety, reliability, and resilience in relation to robustness. The future versions may also cover other aspects such as accuracy, privacy, and validity.

2. Types of data: This version covers structured data, including time series and tabular data, where each row is independent of the other. Additionally, it covers unstructured data in the form of text. Future versions of this standard may cover other types of unstructured data, such as images and speech, as well as various models built upon this data.

3. Types of Models: This standard presents processes for evaluating robustness across all models for structured data and unstructured text data. It encompasses methods for testing open, grey, and closed-box models. Future versions may extend coverage to include other types of models such as Reinforcement Learning, Generative Adversarial Networks (GANs), and Autoencoders.

4. Types of components: The current version covers robustness assessment of data, ML models, and AI systems. Future versions may cover other components such as interfaces, pipelines, infrastructure, and deployments.

5. Type of lifecycle stages: This version covers the data lifecycle, model build lifecycle, and counterfactual deployment scenarios.

6. Types of metrics: While this standard presents a range of robustness metrics and combined metrics, it does not endorse any specific metric. The choice of metrics must be determined on a case-by-case basis, considering the system's nature, domain, and underlying use cases.

7. Vulnerability Mitigation: While the standard presents various strategies for mitigating vulnerabilities, the implementation of mitigation measures falls outside its scope and is left to the developer of the AI system.

2.2. Users of Standard

2.2.1. Organizations/ Individuals developing AI systems

One goal of the Standard is to help the AI developer achieve a set of robustness scores for the AI system under development through self-assessment using the recommended SOPs within the framework. Hence, the first-level user of the report would be the AI system developer.

The second-level user would be the auditor or tester responsible for auditing the AI system. The robustness scores, provided by the developer at the first level, along with

the assessment of the developer's adherence to the framework's SOPs, would provide a baseline for the auditor to proceed with further evaluations.

The third-level users would be the management and key decision-makers. These individuals may include policymakers in the government, regulators from a regulatory agency, civil society members involved in AI robustness or ethics work, lawyers, and business leaders who need to decide whether to release the AI tool into production.

2.2.2. Third-party auditors

Independent third-party auditors, accredited by a certifying agency, may audit the AI systems and issue Robustness Certificates with rating scores based on this standard. The sector regulators could either voluntarily adopt or mandate these certifications. The third-party auditors are also responsible for validating the assumptions and choice of parameters used during the self-certification process by the AI tool developer. The auditors are expected to be a team of domain experts, representatives from legal and regulatory bodies, as well as technology and data experts. They should have sufficient domain knowledge to verify the context-specific choices (of protected attribute/metric/threshold selection) made by the auditee. The auditor may seek access to data or statistical properties of data, model, or metrics from the model if there are concerns regarding proprietary information and related intellectual property. However, the auditor should explicitly document these aspects in the report, along with any specific limitations, to ensure comprehensive certification of the AI system.

2.2.3. Procuring organisations

Many organisations follow a transparent tendering process for procurement of goods and services. These include government departments, public sector undertakings, banks, international bodies like the World Bank, and non-governmental organisations (NGOs). They may include AI-based applications in their future procurements. The services offered by these organisations might impact the lives of millions of citizens. It is, therefore, essential for them to deploy only those AI systems that are proven to be robust.

To benchmark the solutions offered by various bidders during the bidding, robustness certificates based on the standardised assessment process could be required as a qualification criterion. Additionally, these organisations might need expertise to assess whether the delivered AI systems are robust. Therefore, these procuring organisations may request self-certification or third-party certification for robustness based on the Standard Operating Procedures (SOPs) enumerated in this standard.

2.2.4. Sector regulators

In specific verticals where robustness in AI systems is crucial, such as critical infrastructure, telecommunications networks, medical diagnosis applications, self-driving cars, and autonomous aircraft, sector regulators may mandate tolerance levels on relevant, carefully selected metrics. They may specify the minimum robustness rating scores as benchmarks for various industry-specific use cases, including specific tolerance levels, if required.

2.2.5. Start-ups and SMEs

Developers, particularly start-ups and SMEs may get their systems certified for robustness from third-party auditors for broader acceptability of their products.

3.0 Normative References

None

4.0 Terms & Definitions

4.1. Terms defined elsewhere

- **Accuracy** [ISO 25000]: The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use[1].
- **Adaptability/ Functional adaptability** [ISO 25000/ ISO 25059]: Degree to which an AI system can accurately acquire information from data, or the result of previous actions, and use that information in future predictions[1], [2].
- **Availability** [ISO 25010]: Degree to which a system, product or component is operational and accessible when required for use [3].
- **Completeness** [ISO 5259]: The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
- **Compatibility** [ISO 25010]**:** Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions while sharing the same common environment and resources. This characteristic is composed of the following sub-characteristics: [3]
    - **Co-existence** - Degree to which a product can perform its required functions efficiently while sharing a common environment and resources with other products, without detrimental impact on any other product.
    - **Interoperability** - Degree to which a system, product or component can exchange information with other products and mutually use the information that has been exchanged.
- **Confidentiality** [ISO 25010]: Degree to which a product or system ensures that data are accessible only to those authorised to have access [3].
- **Consistency** [ISO 5259]: The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.
- **Data Breach** [ITU-T X.1631]: Compromise of security that leads to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to protected data transmitted, stored, or otherwise processed [5].
- **Data Integrity** [ITU-T X.800]: The property that data has not been altered or destroyed in an unauthorised manner [6].

- **Efficiency** [ISO 5259]: The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.
- **Error** [ISO/IEC 2382-14]: A discrepancy between a computed, observed or measured value or condition and the true, specified or theoretically correct value or condition [7].
- **Error Handling** [ISO 16484-5:2022(en)]: A procedure used to identify the presence of errors in a communication [8].
- **Explainable Machine Learning** [ETSI GS ZSM 012 V1.1.1 (2022-12)]:Machine Learning model that can explain its decisions to humans in a comprehensible manner [9].
- **Fail safe** [ISO 25010]: Degree to which a product can automatically place itself in a safe operating mode, or to revert to a safe condition in the event of a failure [3].
- **Fault tolerance** [ISO 25010]: Degree to which a system, product or component operates as intended despite the presence of hardware or software faults [3].
- **Faultlessness (Maturity)** [ISO 25010]: Degree to which a system, product or component performs specific functions without fault under normal operation [3].
- **Flexibility** [ISO 25010]: Degree to which a product can be adapted to changes in its requirements, contexts of use or system environment. This characteristic is composed of the following sub-characteristics: [3]
    - **Adaptability** - Degree to which a product or system can effectively and efficiently be adapted for or transferred to different hardware, software or other operational or usage environments.
    - **Scalability** - Degree to which a product can handle growing or shrinking workloads or to adapt its capacity to handle variability.
    - **Installability** - Degree of effectiveness and efficiency with which a product or system can be successfully installed and/or uninstalled in a specified environment.
    - **Replaceability** - Degree to which a product can replace another specified software product for the same purpose in the same environment.
- **Functional Suitability** [ISO 25010]: This characteristic represents the degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions. This characteristic is composed of the following sub-characteristics: [3]
    - **Functional completeness** - Degree to which the set of functions covers all the specified tasks and intended users' objectives.
    - **Functional correctness** - Degree to which a product or system provides accurate results when used by intended users.
    - **Functional appropriateness** - Degree to which the functions facilitate the accomplishment of specified tasks and objectives.
- **Hazard warning** [ISO 25010]: Degree to which a product or system provides warnings of unacceptable risks to operations or internal controls so that they can react in sufficient time to sustain safe operations [3].
- **Integrity** [ISO 25010]: Degree to which a system, product or component ensures that the state of its system and data are protected from unauthorised modification or deletion either by malicious action or computer error [3].

- **Interoperability** [ISO/IEC 25010:2023(E)]: Capability of a product to exchange information with other products and mutually use the information that has been exchanged. Note 1 to entry: Information is meaningful data; and information exchange includes transformation of data for exchange [3].
- **Intervenability** [ISO 25060]: Degree to which an operator can intervene in the operation of an AI system in a timely manner to avoid damage or danger [10].
- **Machine learning algorithm** [ISO/IEC 22989]: Algorithm to establish parameters according to a given criteria, of a machine learning model from data [11].
- **Maintainability** [ISO 25010]: This characteristic represents the degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements. This characteristic is composed of the following sub-characteristics: [3]
    - **Modularity** - Degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.
    - **Reusability** - Degree to which a product can be used as an asset in more than one system, or in building other assets.
    - **Analysability** - Degree of effectiveness and efficiency with which it is possible to assess the impact on a product or system of an intended change to one or more of its parts, to diagnose a product for deficiencies or causes of failures, or to identify parts to be modified.
    - **Modifiability** - Degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.
    - **Testability** - Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.
- **Non-repudiation** [ISO 25010]: Degree to which actions or events can be proven to have taken place so that the events or actions cannot be repudiated later [3].
- **Operability** [ISO/IEC 25010:2023(E)]: Operability capability of a product to have functions and attributes that make it easy to operate and control [3].
- **Operational constraint** [ISO 25010]: Degree to which a product or system constrains its operation to within safe parameters or states when encountering operational hazard [3].
- **Performance Efficiency** [ISO 25010]: This characteristic represents the degree to which a product performs its functions within specified time and throughput parameters and is efficient in the use of resources (such as CPU, memory, storage, network devices, energy, materials...) under specified conditions. This characteristic is composed of the following sub-characteristics: [3]
    - **Time behaviour** - Degree to which the response time and throughput rates of a product or system, when performing its functions, meet requirements.
    - **Resource utilisation** - Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

    - **Capacity** - Degree to which the maximum limits of a product or system parameter meet requirements.

- **Portability** [ISO 5259]: The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.
- **Prediction** [ISO/IEC 22989]: Output of a machine learning model when provided with input data [11].
- **Privacy** [ITU-T X.800]: The right of individuals to control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed. Note – Because this term relates to the rights of individuals, it cannot be very precise [6].
- **Quality** [ISO 25010]**:** The quality model is the cornerstone of a product quality evaluation system. The quality model determines which quality characteristics will be taken into account when evaluating the properties of a software product [3].
- **Recoverability** [ISO 25010]: Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system [3].
- **Reliability** [ISO 25010]: Degree to which a system, product or component performs specific functions under specified conditions for a specified period of time. This characteristic is composed of the following su characteristics: [3]
    - **Faultlessness** - Degree to which a system, product or component performs specific functions without fault under normal operation.
    - **Availability** - Degree to which a system, product or component is operational and accessible when required for use.
    - **Fault tolerance** - Degree to which a system, product or component operates as intended despite the presence of hardware or software faults.
    - **Recoverability** - Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.
- **Resilience Testing** [Building automation and control systems (BACS) — Part 5: Data communication protocol]: A procedure used to identify the presence of errors in a communication [8].
- **Resilience** [ISO 22301]: It enables an organisation to have a more effective response and a quicker recovery, thereby reducing any impact on people, products and the organisation's bottom line [12].
- **Resistance** [ISO 25010]: Degree to which the product or system sustains operations while under attack from a malicious actor [3].
- **Reusability** [ISO 25010]: Degree to which a product can be used as an asset in more than one system, or in building other assets [3].
- **Risk identification** [Risk identification]: It is the process of finding, recognizing and recording risk.
- **Robust Machine Learning** [ETSI GS ZSM 012 V1.1.1 (2022-12)]: Machine Learning model that is resilient to adversarial attacks (e.g. data poisoning, model leakage), that can handle unintentional errors (e.g. missing data, data

drift), that have safeguard mechanisms (e.g. fallback to rule-based algorithms) put in place to deal with unexpected outcomes and that are reproducible [9].

- **Safe integration** [ISO 25010]: Degree to which a product can maintain safety during and after integration with one or more components [3].
- **Safety** [ISO/IEC Guide 51:2014, 3.14, modified]: Freedom from unacceptable risk [13].
- **Safety** [ISO 25010]: This characteristic represents the degree to which a product under defined conditions avoids a state in which human life, health, property, or the environment is endangered. This characteristic is composed of the following sub-characteristics: [3]
    - **Operational constraint** - Degree to which a product or system constrains its operation to within safe parameters or states when encountering operational hazard.
    - **Risk identification** - Degree to which a product can identify a course of events or operations that can expose life, property or environment to unacceptable risk.
    - **Fail safe** - Degree to which a product can automatically place itself in a safe operating mode, or to revert to a safe condition in the event of a failure.
    - **Hazard warning** - Degree to which a product or system provides warnings of unacceptable risks to operations or internal controls so that they can react in sufficient time to sustain safe operations.
    - **Safe integration** - Degree to which a product can maintain safety during and after integration with one or more components.
- **Scalability** [ISO 25010]: Degree to which a product can handle growing or shrinking workloads or to adapt its capacity to handle variability [3].
- **Security** [ISO 25010]: Degree to which a product or system defends against attack patterns by malicious actors and protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization. This characteristic is composed of the following sub-characteristics: [3]
    - **Confidentiality** - Degree to which a product or system ensures that data are accessible only to those authorised to have access.
    - **Integrity** - Degree to which a system, product or component ensures that the state of its system and data are protected from unauthorised modification or deletion either by malicious action or computer error.
    - **Non-repudiation** - Degree to which actions or events can be proven to have taken place so that the events or actions cannot be repudiated later.
    - **Accountability** - Degree to which the actions of an entity can be traced uniquely to the entity.
    - **Authenticity** - Degree to which the identity of a subject or resource can be proved to be the one claimed.
    - **Resistance** - Degree to which the product or system sustains operations while under attack from a malicious actor.
- **Security** [ITU-T X.800]: The term "security" is used in the sense of minimising the vulnerabilities of assets and resources. An asset is anything of value. A

vulnerability is any weakness that could be exploited to violate a system or the information it contains. A threat is a potential violation of security [6].

- **Test data** [ISO/IEC 22989]: Data used to assess the performance of a final machine learning model [11].
- **Threat** [ISO/IEC 27000]: Potential cause of an unwanted incident, which can result in harm to a system or organisation [14].
- **Training data** [ISO/IEC 22989]: Subset of input data samples used to train a machine learning model [11].
- **Usability** [ISO/IEC 25010]: Degree to which a product or system can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use [3].
- **Unauthorised access** [ITU-T M.3016.0]: An entity attempts to access data in violation of the security policy in force [15]
- **Validation** [ISO/IEC 22989]: Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled [11].
- **Vulnerability** [NIST-SP-800-30]: A weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source [16].
- **Vulnerability Management** [ITU-T X.1361]: The process that consists of identifying, classifying, remediating, and mitigating vulnerabilities [17].

4.2. Terms defined in this document

Robustness: The degree to which an AI system maintains its functional correctness and remains insensitive to specific adversarial phenomena in the data, model, human-in-the-loop, integration or interfaces or deployment environment, thereby limiting privacy exposures, safety issues or security incidents, reliability or resilience failures, and causal inconsistencies.

5.0 Overview of Robustness

5.1. Robustness in the context of AI

5.1.1. Concept

Robustness in this standard will be explored from the standpoint of the entire AI ecosystem. The AI value chain consists of various processes (procurement, design, development, deployment, and post-market monitoring stages, including data collection, pre-processing, model design, model validation, model deployment, and model monitoring) and components (data, model, pipeline, infrastructure, interface, integrations, deployment environment, and Human-in-the-loop). Robustness in an AI system has the potential of being assured when its entire ecosystem (process and component) demonstrates certain Core Principles and Core elements of AI robustness in its overall deployment and applications. The core principles lay down the 'what' aspect of robustness, while the core elements reflect 'how' aspects of the robustness. The core principles and elements are detailed below.

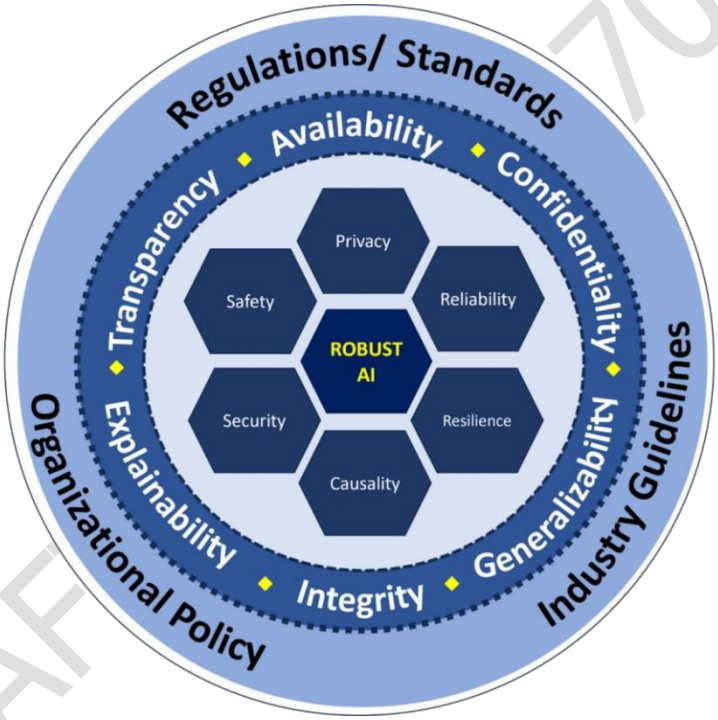

Figure 1: Core Principles and Core Elements of AI

5.1.1.1. Core Principles

Availability: A crucial factor in the robustness of AI is its availability. It refers to the ability of an AI system to be accessible and operational when needed. It emphasises the importance of AI systems to be reliable and dependable, enabling them to consistently perform their intended functions without experiencing disruptions or

downtime. It applies to all components of the value chain of the AI ecosystem requiring them to perform consistently and accurately even in face of unexpected inputs or conditions while maintaining high performance levels over time. To ensure availability of AI systems, it is important that systems are developed on a wide range of inputs and conditions and deploy the system on a reliable and scalable infrastructure to monitor its performance.

Confidentiality : Confidentiality refers to protection of sensitive information and data handled by an AI system. It ensures that the system maintains the privacy and security of user data, preventing unauthorised access, disclosure or misuse. Since AI systems often rely on large datasets to train and operate, data confidentiality becomes a crucial concern for protecting sensitive data that relates to personal information, trade secrets or intellectual property. Developers can ensure robustness by employing encryption techniques, access controls, and data masking to mitigate these risks.

Integrity: Integrity refers to the trustworthiness and correctness of the outputs and behaviour of an AI system. It involves ensuring that the system operates as intended and produces accurate and reliable results. Integrity within AI may be taken care of by strategically mapping and governing the potential noise around the value chain. Its sustenance is reliant on a range of planned interventions such as real time monitoring, enhancing data resilience, alert risk management and active feedback loops. Further, mechanisms for quick error detection, recovery, redressal mechanisms and including human intervention to ensure where AI systems cannot be trusted may be considered.

Transparency: Transparency in AI robustness refers to the ability to understand and explain the reasoning behind the decisions made across the AI lifecycle. It entails making the decision-making process and underlying algorithms of the system open and understandable to users and stakeholders. AI developers and deployers should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

Explainability: Explainability refers to the ability to provide understandable explanations for the decisions and actions taken by an AI system. It involves clarifying why and how the system arrived at a particular outcome, providing insights into its internal processes and decision-making. AI systems allow the processing of large amounts of data, automation of processes as well as the detection of patterns in datasets. Yet the complexity of AI systems, in particular AI systems using ML approaches, may render the evaluation of the results validation as a major challenge. Methods and procedures which ensure that the results presented by AI systems need to be understood and evaluated. In particular, the enhancement of the explainability of the outputs will be crucial to guarantee that disputes between stakeholders in the telecommunications sector can be tackled. AI explainability relates to the means that allow users to understand and trust AI outputs. Further, explainability also becomes a crucial element of robustness as AI models including Deep learning models do not necessarily have adequate interpretability / explainability leading to exposures/ failures.

Generalizability: Generalizability refers to the ability of an AI system to perform well and provide accurate results on data that it has not been trained on. It involves the ability of the system to apply knowledge and insights gained from one dataset to other similar datasets.Generalizability is a crucial principle in robustness in AI as it refers to an AI system's ability to perform well on data it hasn't been trained on. It ensures that the system can handle different scenarios and variations in the real world, builds trust and confidence in its capabilities, and contributes to fairness and ethical aspects.

Achieving generalizability involves using diverse training datasets, employing techniques like cross-validation and transfer learning, and focusing on developing robust AI systems that can effectively handle new and unseen data.

5.1.1.2. Core Elements

Robustness is a critical aspect in the development and deployment of AI systems, and it encompasses several core elements. These elements include privacy, reliability, safety, security, resilience, and causality. Privacy ensures the safeguarding of user data, preventing unauthorised disclosure or misuse. Reliability focuses on consistently producing accurate and trustworthy outcomes, while safety involves protecting users and the environment from potential harm. Security is vital for maintaining the integrity and confidentiality of AI systems, guarding against adversarial attacks and vulnerabilities. Resilience enables the system to recover and adapt in the face of disruptions, ensuring continuity of operations. Further, causality plays a crucial role in understanding the cause-and-effect relationships within AI systems, aiding in the identification and prevention of potential risks/ vulnerabilities.

Privacy
In order to safeguard user data and prevent the disclosure, leakage, or improper use of sensitive information, privacy is a fundamental component of the robustness of AI systems. AI systems must handle and process data securely, protect user privacy, and adhere to legal and ethical standards in order to be robust. Privacy exposure and leakage are challenges that arise in robustness discussions. Instances of private data being inadvertently disclosed or compromised during data processing or model training are examples of such leakage and exposure. Adversarial attacks, including jailbreaks and prompt attacks, introduce additional intricacy by deliberately manipulating AI systems in order to exploit vulnerabilities; this may have adverse effects on user privacy.

Reliability
Reliability is crucial for building trust in AI systems and ensuring accurate decision-making. Robustness requires AI systems to consistently produce reliable outcomes, responses, and predictions. When users rely on the outputs of AI models, they need confidence in their reliability to make informed decisions and take appropriate actions. Relying on model outcomes in AI systems presents challenges such as uncertainty and biases. Uncertainty arises when models struggle with ambiguous situations, leading to unpredictable outcomes and posing difficulties in critical decision-making. While research is expanding on approaches to quantifying uncertainty, it's important to consider that as a factor in robustness, as uncertainty may lead to incorrect outcomes/ predictions. Biases and fairness issues can arise when models exhibit unreliable outcomes and unfair decisions. Additionally, reliability of AI systems can lead to confusion, distrust, and flawed decision-making.

Safety
Safety is an essential consideration in the context of robustness in AI. It involves protecting users and the environment from potential harm caused by dangerous suggestions or incorrect predictions. Specifically in use case environments where the outcomes/ predictions could lead to significant impact to people (e.g. predicting offenders) or planet. Challenges arise when models provide harmful recommendations or when validating the accuracy of predictions is complex due to subjective evaluation. To ensure robustness, ongoing validation is necessary with evolving data that reflects real-world scenarios.

Security
The relevance of security as a core element in the robustness discussion is crucial to ensure the integrity, availability, and confidentiality of AI systems. Cybersecurity challenges and adversarial exposure pose significant risks to the robustness of AI systems. Adversarial attacks involve deliberately manipulating input data to deceive or exploit AI models, leading to incorrect or malicious outputs. Additionally, vulnerabilities in data, models, deployments, human-in-the-loop processes, and integrations/interfaces can introduce complexity and potential security risks. While approaches including AI red teaming or bug bounty programs are evolving, there needs to be more structured approaches to managing security of AI systems.

Resilience
Resilience is a critical element in robustness as it focuses on an organisation's ability to recover and adapt in the face of disruptions or attacks. In the context of AI systems, resilience involves the ability to recover from failures, attacks, or other adverse events, ensuring the continuity of operations. Enterprise disaster planning plays a crucial role in achieving resilience by developing strategies, contingency plans, and recovery mechanisms to mitigate the impact of disruptions. Lack of resilience in a distributed environment (The data pipeline, the interface, the model and the applications operate in multiple layers of infrastructure), where AI models are integrated with multiple applications, can result in severe consequences and significant resource loss.

Causality
Causality is a crucial element in robustness as it focuses on understanding the cause-and-effect relationships within AI systems. Analysing causality helps identify potential vulnerabilities and risks, allowing organisations to take proactive measures to prevent disasters and protect people. However, there are challenges associated with causal analysis, particularly in high-risk systems. One challenge is the limitation of causal analysis techniques, which may not capture complex causal relationships accurately. Additionally, the lack of adequate causal information in the data can hinder the ability to identify and mitigate potential risks.

5.1.2. Requirements regarding AI robustness

The requirements regarding AI robustness shall be compiled from 3 key sources, namely, the regulatory requirements, the industry requirements and the enterprise / company policy requirements. Many such requirements are evolving towards a risk based approach to evaluate robustness of AI systems. Risk based approach enables organisations to approach robustness based on the priority and impact of the risk, thereby allowing the organisation to focus on aspects that can scale effectively and efficiently.

Although India does not have comprehensive regulations pertaining to robustness, legislation such as the EU AI Act offers a more comprehensive outline of the necessary criteria concerning robustness. As an illustration, the European Union Artificial Intelligence Act (recital 50, dated February 6th, 2024) specifies that inadequate robustness may give rise to safety implications and/or erroneous decisions or biased outputs that have adverse effects on fundamental rights. A summary of expectations from Recital 50 are provided below:

Topics	Expectations

General	• The ***technical robustness*** is an essential requirement for high-risk AI systems. • They should be ***resilient*** to harmful or otherwise undesirable behaviour that may have limitations within the systems or the environment.
Organisational measures	• ***Technical and organisational measures*** should be taken to ensure robustness.
Specific expectations	• Appropriate technical solutions should be designed and developed. • These solutions aim to ***prevent or minimise harmful*** or otherwise undesirable behaviour. • Mechanisms enabling ***safe interruption of system operation*** may be included. • Fail-safe plans are essential when ***anomalies are detected***, or predetermined boundaries are exceeded.

This standard aims to bring the industry requirements regarding AI robustness. Organisations may have specific policies and requirements regarding AI robustness in their operations.

5.2. Robustness in the context of AI in telecom and digital infrastructure

The Indian telecom industry stands as the world's second-largest, showcasing robust growth and significant trends [18], [19]. Notably, as of September 2023, the industry's overall tele-density is ~85%, with the rural market presenting untapped potential at 58%, while the urban tele-density stands at 134%, underlining the dynamic landscape and vast opportunities within the Indian telecom sector [19].

The rural market, encompassing 70% of the population, emerges as a key growth driver. The sector boasts an impressive 116 crore mobile connections, with 70 crore internet users and 60 crore smartphone users, reflecting widespread technological adoption. With smartphones averaging 9.8 GB of monthly data usage, the country is at the forefront of digital connectivity [18]. Looking ahead, India anticipates 88 million 5G connections by 2025 [20]. The inherent capabilities of AI, coupled with the extensive telecom infrastructure, are expected to empower diverse segments of the population, further accelerating digital inclusion and technological accessibility across the country.

5.2.1. AI usage in network management

AI is revolutionising telecom network management by bringing proactive maintenance, network optimization, and self-healing capabilities. On the proactive side, AI can analyse network data to predict and prevent problems before they occur. This includes identifying unusual activity that might signal security threats and taking steps to mitigate them. Additionally, AI can analyse equipment data to anticipate potential failures, allowing for preventative maintenance and minimising downtime. AI also optimises networks and manages traffic. It automates network functions based on real-time data, leading to greater efficiency and flexibility. This includes tasks like smart

channel management, energy-efficient network configuration, and user-centric network optimization that prioritises bandwidth for critical services. AI can even analyse usage patterns to predict future demand and plan network upgrades more effectively. Furthermore, AI can automate self-healing capabilities in telecom networks. By autonomously identifying and resolving issues, AI minimises downtime and ensures faster service restoration. This can involve restarting malfunctioning cell sites or optimising resource allocation based on real-time data [21], [22], [23].

Area	Use-cases
Proactive Network Management with AI	• Predictive Maintenance • Anomaly Detection • Proactive Fault Detection and Resolution • Security Threat Detection and Prevention
Network Optimization and Traffic Management, Network Planning and Resource Allocation, Network Automation and optimization management	• Automating Network Management • Safeguarding 5G Networks • Smart Channel Management • Energy-Efficient Networks • Intelligent Cell Clustering • Signal Processing and Spectrum Management • Self-Learning Femtocells (a small, low-power cellular base station) • User-Centric Network Optimization • Predictive Network Planning • Network Optimization Across Stages • Investment-Focused Network Analysis • Targeted Network Improvement • Dynamic Network Adjustments • Enhanced Network Efficiency and Customer Satisfaction

Automated Self-Healing Networks	• Self-Planning
	  ○ Planning location of a new node
	  ○ Planning radio and transport parameters of a new node
	  ○ Planning data alignment for all neighbour nodes
	• Self-Optimization
	  ○ Support for a centralised optimization entity
	  ○ Interference control
	  ○ QoS-related parameters optimization, load balancing
	  ○ Transport parameters optimization, routing optimization

	○ Energy saving • Self-Deployment ○ H/W installation ○ Transmission setup ○ Node authentication ○ Automatic inventory ○ Self-test • Self-Healing ○ H/W capacity expansion / replacement ○ S/W upgrade ○ Network monitoring such as cell / service outage detection, and information correlation for fault management • Failure recovery such as cell outage compensation, and mitigation of unit outage
Energy Efficiency Optimization	• Network energy consumption patterns analysis

Applicability of robustness of AI

The robustness of AI plays a critical role in ensuring the success of the functions outlined for telecom network management.

- **Anomaly Detection:** This relates to Availability. Here, robustness in AI refers to its ability to maintain a low false alarm rate. In telecom networks, overly sensitive AI could mistake normal traffic fluctuations for cyberattacks, triggering unnecessary mitigation actions and disrupting service. Robust AI can effectively distinguish anomalies from regular patterns, minimising false positives and ensuring timely intervention for genuine threats.
- **Generalizability:** For network optimization tasks like traffic management and resource allocation, robust AI needs to generalise well from the training data to unseen scenarios. Imagine an AI trained on traffic patterns in a city being deployed in a rural area. Robust AI should be adaptable enough to learn these new patterns and optimise the network effectively despite the difference in usage.
- **Data Quality and Bias:** The effectiveness of AI for predictive maintenance and network planning hinges on the quality and representativeness of the training data. Robust AI algorithms are less susceptible to biases in the data, leading to more accurate predictions. For instance, biased data on cell tower usage could lead to under-investment in rural areas. Robust AI can help mitigate such biases and ensure optimal network planning across diverse regions.
- **Explainability and Transparency:** In critical areas like self-healing networks where AI autonomously resolves issues, explainability is crucial. Robust AI should be able to provide clear reasoning behind its decisions, allowing network operators to understand the root cause of problems and validate the AI's actions. This transparency builds trust in AI-driven network management.
- **Continuous Learning and Adaptation**: Telecom networks are dynamic environments with evolving traffic patterns and security threats. Robust AI should continuously learn and adapt to these changes. This could involve retraining models with new data or incorporating online learning algorithms that update the

AI in real-time. By continuously adapting, AI ensures the network remains optimist and secure over time.

5.2.2. GenAI/ LLMs in Telecom Use-cases

Large Language Models (LLMs) offer significant value in various downstream tasks within the telecom industry and research is exploring industry specific language models [24]. The use cases relate to transformative impact on personalised experiences, network optimization, customer support, automated operations, and new product/service development [25], [26], [27].

Area	Use-cases
Customer Service	Customer-facing chatbots, Call-routing performance, Agent copilots and Bespoke invoice creation
Marketing and Sales	Content generation, Hyper-personalization, Copilots for store personnel and Customer sentiment analysis
Network	Network inventory mapping, Network optimization via customer sentiment analysis and Enabling self-healing via customer sentiment analysis on network problems
IT	Copilots for software development, Synthetic data generation, Code migration, IT support chatbots, Analysing technical documents, generating reports and presentations and Automating tasks like network troubleshooting.
Other Functions	Procurement optimization, Workplace productivity, Internal knowledge management, HR Q&A, New product and service development

Applicability of robustness of AI

The robustness of AI plays a critical role in ensuring the success of functions such as customer service, sales and marketing, IT and others in telecom.

- **Data Quality and Bias**: For functions like AI chatbots and personalised marketing, the quality and representativeness of training data are paramount. Robust AI algorithms are less susceptible to biases in the data, leading to more accurate and fair outcomes. Biased data on customer preferences could lead to unfair promotions or exclusion of certain demographics. Robust AI mitigates such biases and ensures all customers receive a positive experience.

- **Explainability and Transparency**: In areas like IT operations where AI copilot systems suggest code or identify bugs, explainability is vital. Robust AI should be able to provide clear reasoning behind its suggestions, allowing developers to understand the logic and make informed decisions. This transparency builds trust in AI-powered tools and fosters collaboration between humans and AI.

- **Generalizability**: GenAI specifically needs to generalise well from customer sentiment data to identify network issues. Imagine an AI trained on sentiment data from a mostly urban network being deployed in a rural area. Robust GenAI should be adaptable enough to learn the new patterns of customer sentiment regarding network problems and pinpoint areas requiring attention in diverse network environments. This ensures efficient maintenance efforts and avoids biases towards specific regions.

- **Continuous Learning and Adaptation**: Customer preferences, network usage patterns, and security threats are constantly evolving. Robust AI should continuously learn and adapt to these changes. This could involve retraining models with new data or incorporating online learning algorithms that update the AI in real-time. By continuously adapting, AI ensures its recommendations and actions remain relevant and effective over time.

- **Accuracy and Reliability**: Across all departments, robust AI translates to accurate and reliable outputs. In tasks like procurement optimization or IT support chatbots, even minor errors can have significant consequences. Robust AI minimises errors through rigorous testing and validation, ensuring its recommendations and actions are trustworthy and reliable.

5.2.3. AI use-cases in futuristic telecom scenarios

Area	Use-cases
6G	• **AI-powered network slicing:** AI can dynamically allocate network resources based on real-time needs. Imagine a self-driving car needing ultra-low latency for critical manoeuvres, while a remote surgery requires high bandwidth for data transmission. AI can carve virtual networks (slices) to cater to these diverse demands efficiently [28].
Autonomous Vehicles	• **Predictive maintenance for vehicles:** AI can analyse sensor data from autonomous vehicles to anticipate part failures and schedule maintenance pro-actively. This reduces downtime and ensures safety on the road. • **Real-time traffic management:** AI can analyse traffic patterns and predict congestion. It can then optimise traffic flow by rerouting vehicles and providing real-time information to drivers, reducing travel times and accidents [29].
V2V Communication for OTA (Over-the-Air) Updates	• **Secure and efficient data exchange:** AI can secure communication between vehicles (V2V) for software updates and data sharing. This ensures vehicles receive the latest updates without needing physical intervention, improving overall safety and performance [30].

Applicability of robustness of AI

The robustness of AI plays a critical role in these futuristic scenarios.

Digital Twins in Manufacturing	• **AI-powered anomaly detection:** AI can analyse data from a physical production line's digital twin to identify potential issues before they occur. This proactive approach keeps production lines running smoothly and minimises downtime. • **Predictive maintenance for machinery:** Similar to vehicles, AI can analyse sensor data from the digital twin to predict equipment failures and schedule maintenance. This optimises production schedules and reduces costs [31].

- **Accuracy and Reliability**: In a network with diverse needs (6G network splicing), accurate AI is crucial. Incorrect resource allocation could lead to delays in critical surgeries or accidents for self-driving cars. Robust AI ensures precise network slicing, delivering the promised speed and reliability for each application.

- **Data Quality**: Predictive maintenance relies heavily on sensor data quality. Biased or faulty data could lead to missed warnings about critical part failures. Robust AI minimises the impact of such issues by filtering and analysing data effectively, ensuring accurate predictions.

- **Generalizability**: In case of self-driving cars, traffic patterns can vary significantly between cities and rural areas. Robust AI should be adaptable enough to learn from new sensor data and adjust predictions accordingly. This ensures reliable maintenance recommendations regardless of location.

- **Security and Privacy**: Secure communication is paramount for V2V updates. Robust AI should be resistant to hacking attempts and data breaches. This safeguards vehicles from malicious software and protects user privacy.

- **Explainability and Transparency**: When AI identifies potential issues in the digital twin, manufacturers need to understand the reasoning. Robust AI should provide clear explanations behind its predictions, allowing engineers to verify the issue and take appropriate action.

- **Continuous Learning and Adaptation**: Manufacturing processes and equipment can evolve over time. Robust AI should continuously learn from new data collected by the digital twin. This ensures accurate anomaly detection and adapts to changes in the production line.

5.2.4. Incidents due to lack of robustness of AI in other sectors

The robustness of AI Systems may get compromised owing to multiple reasons. The system may receive abnormal or unexpected inputs that may lead to malfunction. Further, the systems attempt to achieve a different outcome from what the designer/ operator intended, may result in unexpected behaviours or side effects. Inadequate monitoring of the operations given the opacity of the neural networks may also compromise upon the robustness of AI systems.

Some of the prominent cases where the robustness of an AI system was challenged have been provided below:

- **Cancer detector misdiagnoses black users**: AI systems may fail when applied to circumstances which deviate from their intended purpose or where inputs aren't identical to those used during training. This was observed in the case of America, where a self-screening software was developed to diagnose early-stage indicators of skin cancer on the phone. Millions of Americans used the app to diagnose symptoms and a few years later, public health experts claimed that the detector misdiagnosed the black users. The app diagnosed a dramatic increase of late-stage skin cancer among Black patients' and post an investigation, it was concluded that the self-screening software is significantly less accurate at identifying malignancies on people with dark skin tones as the training mostly represented northern Europe.
- **Bus ad triggers facial recognition system (FRT)**: IntelliMotor designed an AI-based vision system for its new driverless iTaxis to identify human faces near the windscreen. The feature helped in ensuring public safety and fostering confidence in the technology by automatically slowing down when detecting a human face with a high degree of certainty. The iTaxis next received a software upgrade with the introduction of a new facial recognition feature. The face recognition feature of the iTaxis was however triggered when advertisements for a concert (consisting of human faces) were put on several city buses. The FRT software recognized every face on the advertisement on the bus and instantly stopped, causing multiple random stops whenever approached by buses, thus resulting in thousands of crashes across the nation.
- **Absence of subjective interpretation**: An incident occurred where the navigation apps detected low traffic on nearby side roads and began redirecting drivers accordingly to the empty lane. The application, however, did not consider that the roads were empty because the surrounding neighbourhoods were evacuated due to fire in the nearby area. Hence, the apps 'algorithm did not consider fire safety conditions and directed the traffic to the side roads. Resultantly, when the wind picked up, the wildfire quickly spread into the evacuated area, thereby trapping the rerouted vehicles in the flames.
- **AI fails on the high seas**: The Morsen Shipping Lines implements a sophisticated computer vision system. The system can identify obstacles and approaching vessels with high speed and precision in low visibility conditions. However, an instance was encountered where, when the tanker approached a semi-submerged trash off the coast of Florida, the vision system failed to sound a warning for reasons which Morsen's technical experts are still trying to figure out. Resultantly, carcinogenic substances leaked out of the ship's hull caused by the debris.
- **Ambulance chaos**: When an exceptionally severe flu season results in a spike in ER visits, the hospitals in New York City resort to using the machine learning platform Routr. Routr uses real-time data reading from member hospitals, public health organisations, and first responders to reroute incoming 911 calls from hospitals that may shortly reach capacity to those that are likely to have sufficient space. Based on AI algorithms that have been trained on terabytes of historical occupancy data, the programme is able to recognise trends that would have been impossible for a human to notice. Thanks to Routr, city hospitals have extra beds in November and December despite a sharp increase in patients. But on New Year's Day, January 1, the software strangely started forwarding calls from all over the city to a select few Queens hospitals. By dawn, the hospitals were overflowing, patients were suffering and, in some cases, dying in traffic jams inside ambulances outside hospital entrances. A state-ordered examination conducted to detect the malfunction concluded that human dispatchers keeping an eye on Routr were aware of the odd routing

pattern but did nothing about it as they were unsure of the details and assumed the AI knew what it was doing.

5.3. Sources of robustness risks

5.3.1. Understanding the sources of Robustness risks

Robustness risks typically arise from risks contributed by the process or components on one side and risks contributed by inadequacies in mitigations on the other side.

- Process refers to the design, development, deployment, and post-market monitoring stages, including data collection, pre-processing, model design, model validation, model deployment, and model monitoring.
- Component refers to the data, model, pipeline, infrastructure, interface, integrations, deployment environment, and Human-in-the-loop.
- Mitigations-related inadequacies relate to security governance, design, implementation, verification, and operations.

The summary of key robustness risk contributed by process and components are listed below.

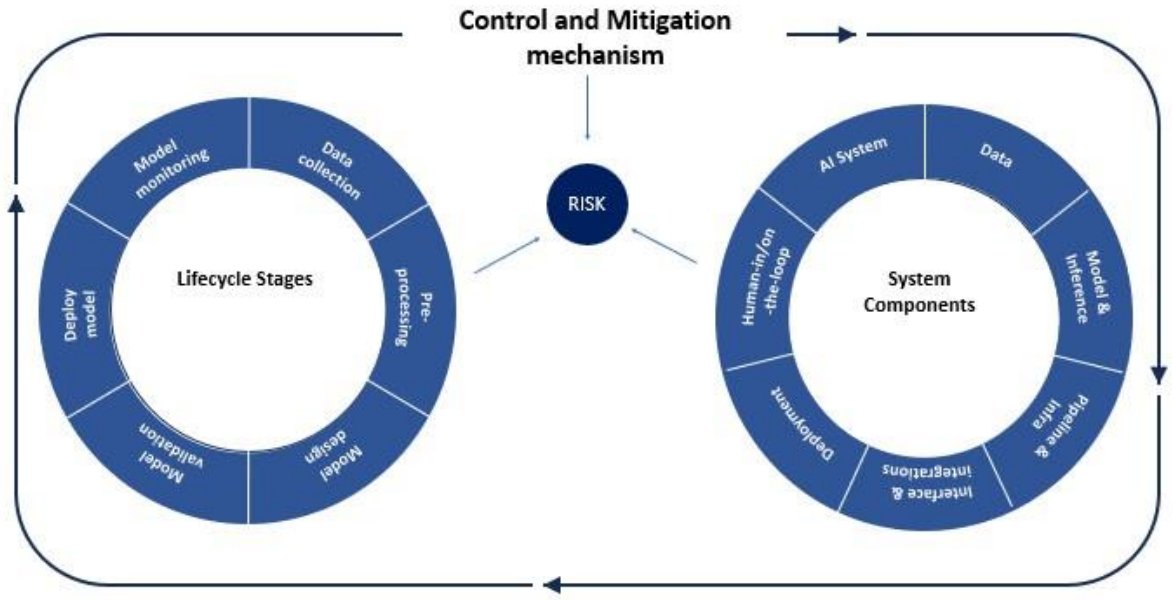

Figure 2: Sources of robustness risks

5.3.1.1. Robustness risk contributed by process/ Lifecycle stages

Process	Robustness risk contributors
Data collection	• Inaccurate data collection methods.

	• Lack of standardised procedures for data collection. • Insufficient training for personnel involved in data collection. • Data collection sensors, tools, or equipment. malfunctioning, affecting the quality of data. • Inadequate data validation processes. • Security breaches or data loss during the data collection process. • Failure to account for potential biases in data collection. • Inadequate documentation of data collection processes.
Pre-processing	• Exploitative data imputations leading to backdoor attacks [32]. • Inconsistent outlier detection and handling. • Data or batch normalisation and scaling strategies exposing vulnerabilities [33]. • Insufficient cross validation of the data. • Improper database connection closures.
Model development	• Inconsistent parameter tuning. • Inappropriate optimization strategies. • Code complexity or non-modularity. • Unvalidated compilers [34]. • Insufficient error handling. • Lack of adequate redundancy mechanism.
Model validation	• Inadequate validation of results [36]. • Insufficient security validation checks. • Lack of mechanism to conduct scenario testing performance for deployment environment.
Model deployment	• Lack of resources for scalability. • Inadequate documentation on model deployment.
Model monitoring	• Insufficient error handling. • Inadequate logs or metrics for performance and/ or quality. • Lack of mechanism for timely calibration of tools and sensors. • Inadequate feedback mechanism for model enhancement. • Lack of mechanism to track threats or adverse incidents.

5.3.1.2. **Robustness contributed by components**

Components	Robustness risk contributors
Data	• Insufficient data quality. • Lack of mechanism to monitor data drifts. • Inadequate mechanism to validate user inputs to prevent vulnerabilities like injection attacks. • Inadequate data protection measures. • Backdoors threats arising from data sources.
Model	• Model sensitivity to input perturbations or diversity. • Inadequate adversarial preparedness of the model. • Lack of mechanism to monitor model drifts. • Insecure coding practices contributing to vulnerabilities during software development.

	• Vulnerabilities contributed by open-source or sourced models.
Pipeline	• Inadequate fault tolerance. • Lack of monitoring mechanism for resource usage. • Lack of mechanism to track failed data validation or transformation.
Infrastructure	• Inadequate implementation of redundancy systems or backup management. • Insufficient mechanism to handle cloud security. • Over reliance on specific infra.
Interface	• Under prioritised security and safety. • Inadequate bug tracking of interfaces.
Integration	• Inadequate testing of integrations. • Performance degradation due to integration. • Insufficient controls over version management.
Deployment environment	• Inadequate logging and monitoring mechanisms to detect and respond to security incidents promptly. • Insufficient documentation regarding the metrics, methods, and thresholds for monitoring.
Human-in-the-loop	• Inadequate expertise to examine or observe the triggers, threats, errors, or omissions. • Insufficient user and rights management.
AI System as a whole	• Inadequate interpretability of outcomes or model operations. • Improper error handling mechanisms to handle unexpected situations and prevent system crashes. • Lack of mechanism for strong encryption mechanisms to protect sensitive data from unauthorised access. • Inadequate mechanism to manage or update patches. • Insufficient access control measures to prevent unauthorised access to sensitive data. • Inadequate disaster recovery planning for data integrity failures or system outages.

5.3.1.3. Robustness risk contributed by inadequate mitigation

Inadequacies in mitigations shall also be a contributor to the robustness risks. Illustrative list of these inadequacies are provided below aligned to the Software Assurance Maturity Framework of OWASP:

Framework element	Topic	Robustness risk contributors
Governance	Strategy and metrics	• Lack of alignment between robustness strategy and organisational goals. • Lack of executive buy-in and support for robustness governance initiatives. • Inconsistent application of robustness governance across different processes in AI lifecycle.

		• Inadequate communication and collaboration between technology teams and business teams on robustness. • Inadequate measures or metrics, or failure to regularly review/ update robustness governance metrics. • Overreliance on outdated or irrelevant robustness measures or metrics. • Insufficient resources allocated to implement and monitor robustness measures or metrics. • Failure to notice or address emerging robustness threats or failure mode. • Inability to adapt robustness metrics and measures to changing regulatory requirements.
	Policy and compliance	• Poorly defined policies may create confusion and compliance issues. • Lack of encryption measures, weak authentication, and insufficient disaster recovery measures.
	Education and guidance	• Inadequate training on robustness best practices to help users identify and spot systems at risk. • Insufficient user training increases susceptibility to social engineering tactics.
Robustness Design	Threat assessment	• Lack of regular robustness assessments. • Inadequate threat modelling. • Failure to monitor network traffic. • Inadequate incident response plan. • Lack of data protection and encryption measures for data in and out of AI systems.
	Robustness requirements	• Insufficient or non-diverse test data. • Inadequate measures to compile essential security requirements.
	Robustness architecture	• Inadequate encryption protocols. • Lack of regular security updates. • Insufficient access controls. • Weak password policies. • Failure to deploy mechanisms to monitor and detect security incidents. • Inadequate network segmentation.
Robustness implementation	Robustness Build	• Vulnerabilities in third-party libraries. • Lack of secure coding practices. • Over-reliance on third-party tools without adequate security or safety diligence. • Failure to regularly update security patches.

	Secure deployment	• Insufficient robustness testing at the time of deployment. • Inadequate security controls. • Failure to update security patches. • Weak encryption protocols exposing sensitive data during deployment. •
	Defect Management	• Lack of thorough security verification. • Failure to address security flaws. • Inadequate defect, failure, or incident management.
Robustness verification	Architecture assessment	• Inadequate measures to periodically review the architecture for robustness. • Inadequate measures to manage feedback mechanisms on failure modes.
	Requirements-driven testing	• Poorly defined security requirements. • Inadequate monitoring mechanism to detect security incidents promptly. • Over-reliance on metrics and automated tools for robustness testing can miss nuanced security threats. • Insensitive handling of user data.
	Robustness testing	• Insufficient testing coverage. • Lack of adequate documentation for interpretation of metrics and testing results. • Failure to calibrate or update testing tools or code. • Lack of skilled personnel for security testing. • Insufficient consideration of real-world attack scenarios. • Failure to address false positives/negatives. • Inadequate monitoring of testing processes.
Robustness Operations	Incident management	• Lack of clear escalation procedures for incident management. • There are insufficient monitoring tools for security incidents. • Failure to update incident response plans. • Inadequate communication channels for coordination among response teams. • Lack of proper documentation on post-incident analysis. • Insufficient resources are limiting incident management.

		• Failure to conduct regular security drills. • Inadequate incident reporting mechanisms.
	Environment management	• Inadequate data security • Inadequate system monitoring • weak physical security • Incomplete patch management • Lack of or inadequate access controls • Third party integration risks • Inadequate alignment with security standards.
	Operations management	• Insufficient backup and recovery procedures. • Lack of regular security audits.

6.0 Metrics Associated with Robustness

Measuring robustness of AI systems requires defining appropriate metrics, benchmarks, and evaluation protocols that capture the desired elements and principles of robustness mentioned earlier. In this chapter, we review some of the existing metrics and methods for assessing robustness, as well as some of the challenges and limitations of these approaches. We also provide some guidance and recommendations for choosing and applying robustness metrics in practice.

6.1. Resilience and Robustness to data shift

Two types of data shift are relevant to robustness - data drift and prediction drift. Data drift refers to changes in the input data distribution, which might affect the model's accuracy or relevance. Prediction drift refers to changes in the output data distribution, which might indicate a shift in the target variable or the model's behaviour.
Data drift can be measured by using a reference dataset and comparing against the current dataset to evaluate if there is a shift between the data distributions. Some of the methods for measuring this include comparing summary statistics, running hypothesis testing, using distance-based methods and simple rule-based checks.

Prediction drift can be measured by examining the distribution of predicted scores, classes or values. A significant shift from prior values can indicate that there is shift in the model behaviour due to data drift or some other reason, which is affecting the model's predictions.
Model resilience refers to its ability to perform well on a wide range of data sets, over a long period of time. Such models will not overfit on a particular data set or the resilience of a model can be measured using the following across multiple runs or cross-validations - smaller standard deviations, less discrepancy between test and validation set errors, consistency of error rates over time, and with input and output drift.

6.2. Integrity

One of the causes of input data shift can be due to corruption or quality issues in input data used either during model training or inference, which can be detected by monitoring the following metrics:

Missing data, corrupted schema and features: Data that is missing, in the incorrect format and in the wrong range for a particular feature can lead to poor or inconsistent data

Outliers: Data statistics and anomaly detection techniques can be used to detect data points that are significantly different from the rest of the data. Special treatment can then be given to these outliers by either removing them, treating them differently from other inputs, or monitoring their frequency to check for changes in the dataset distribution.

6.3. Reliability

Reliability in machine learning models is in the context of adversarial attacks, which can manipulate input data to evade detection. In an adversarial attack, attackers modify input features to mislead the classifier into assigning an incorrect class value. Sometimes, model transparency can also be used to exploit models, as attackers can exploit decision rules revealed by models for designing adversarial attacks. The reliability of a machine learning model can be measured by its robustness against adversarial attacks.

In the context of generative AI, a reliable Large Language Model (LLM) is characterised by its ability to produce outputs that are both informative and factually accurate. A reliable LLM needs to address the following challenges: Misinformation, Hallucination, Inconsistency and Miscalibration. Misinformation and Hallucination can be measured using specialised benchmarks. Inconsistency can be measured by querying the LLM multiple times on the same input and measuring deviations in output.

6.4. Explainability and Transparency

A transparent Machine Learning Model Model supplies reasoning chains and justifications for its predictions or output. Explainability leads to an understanding of the machine learning model and its results or outputs, which can be important for human users, who want to know how the system made a decision, especially for complicated or critical applications. Model explainability is often challenging for "black-box" models, due to the lack of transparency about the model and its decision making process. Some Machine Learning techniques are inherently more explainable than others. For example, linear regression, logistic regression and decision trees provide weights of features used for predictions, which makes them interpretable.

Several techniques exist to measure and improve explainability of models:

1. **LIME:** Local interpretable model-agnostic explanation (LIME): This technique uses a locally interpretable smaller model that reproduces the behaviour of the larger model.
2. **Shapley additive explanations (SHAP):** SHAP calculates the importance of a particular value for prediction. The SHAP value for a particular feature is the average marginal contribution provided to the model across all possible feature combinations.
3. **Permutation feature importance:** In this technique, the input features are perturbed to check which ones cause the largest changes to output predictions when modified
4. **Global surrogate models:** This technique involves creating a surrogate model that behaves similarly to the original model, using loss functions such as KL divergence. The surrogate model is designed to be explainable and can be used to explain the behaviour of the original model.
5. **Saliency mapping:** A saliency map highlights network activation or attention regions, which can help determine the features that contribute to predictions.

6.5. Privacy and Security

Privacy risks occur in Machine Learning Models when they memorise and sometimes regurgitate sensitive information present in training data. ML model privacy is measured using membership inference attack, in which we check if it is possible to

determine if a particular input data sample was used to train the model or not. If the inference attack is successful, it poses a privacy risk. Traditional methods to perform membership inference attacks include training custom classifiers, however, recent metrics such as the privacy risk score overcome many of their limitations, providing fine-grained understanding of privacy risks. In the context of Generative AI, recent studies have shown that such models can exhibit privacy risks using "Jailbreaks", a technique used to prompt a model to reveal sensitive or unsafe information.
When running ML models on sensitive data, it may be necessary to run inference on encrypted data, for which several privacy-preserving ML techniques have been proposed. Membership inference attacks can be used to measure the security of the ML model in such cases.

To effectively implement and measure robustness in AI systems, various tools and libraries are available, each serving distinct functions across different robustness dimensions. The annexure-I provides a categorized list of essential tools that aid in evaluating key robustness metrics, including resilience to data shifts, integrity, reliability, explainability, privacy, and security.

7.0 Proposed Assessment Framework

7.1. Introduction to the Robustness Assessment Framework for AI System:

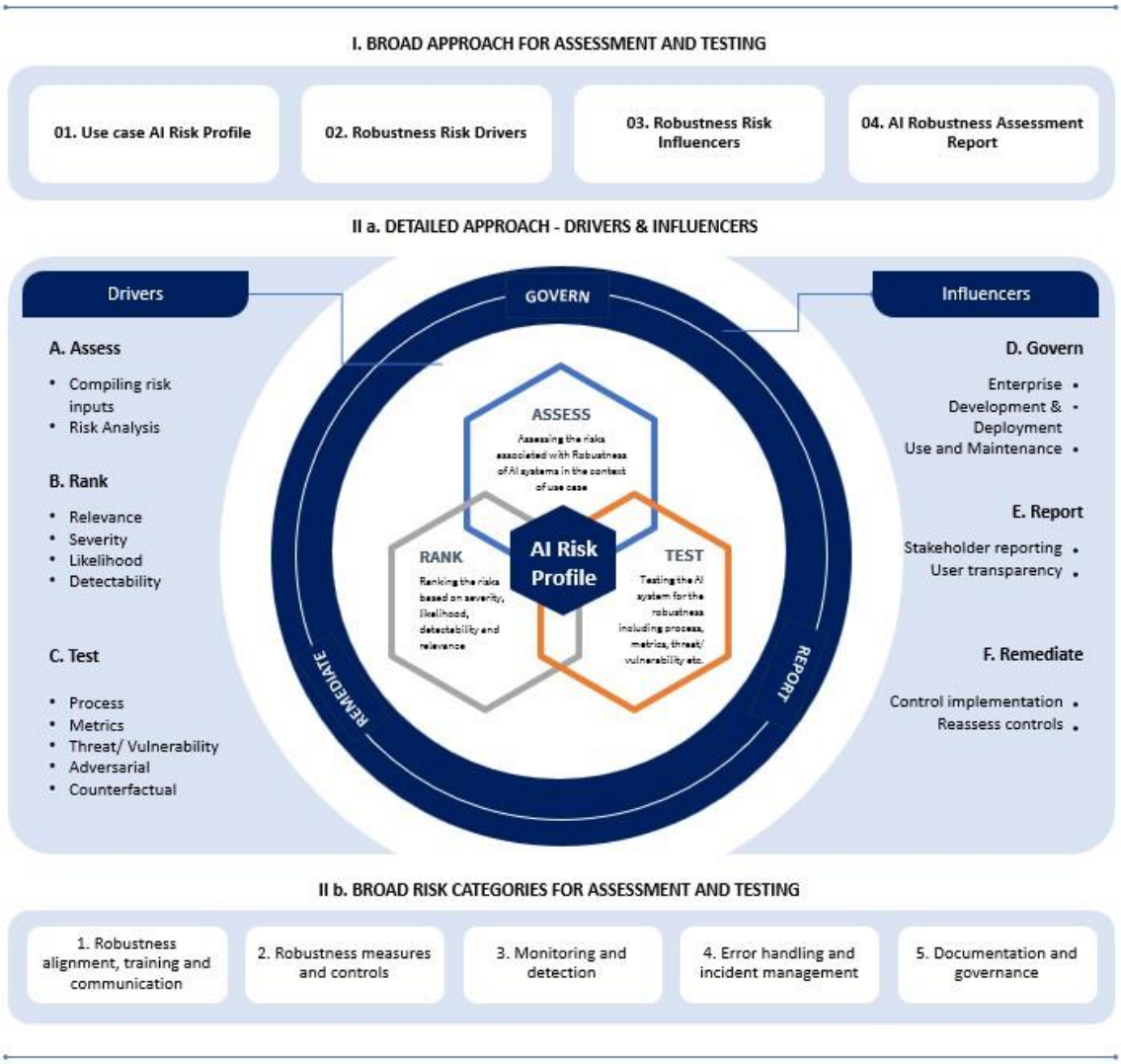

Figure 3: Robustness assessment framework

The robustness assessment framework for AI systems consists of 4 layers, namely, The risk profile layer, the driver layer, the influencer layer and the Ai robustness assessment report layer. These aim to comprehensively assess an AI system's robustness.

The risk profile layer helps in defining the key risks associated with the use case / the AI system that is being examined. The driver layer allows systemic assessment, ranking and testing of the risks associated with robustness in the AI system. The

influencer layer supports the governance, reporting, and remediation efforts with specific reference to the AI system in question. The AI robustness assessment report layer expresses the considerations associated with reporting regarding robustness assessment in conformity with this standard.

By employing these layers, the framework assists in triaging and prioritising risks associated with the AI system. It is important to note that this framework is designed explicitly to conduct robustness assessments of AI systems rather than the overall enterprise.
The framework begins with a focus on the specific use case, allowing for a thorough evaluation of potential risks related to robustness in the drivers layer and weighs in the implications of gaps contributed by the influencer layer. This approach ensures that the assessment addresses the specific challenges and vulnerabilities that may arise within the AI system, enhancing its overall robustness.

7.1.1. Robustness risk profile

Robustness risk profile refers to the process of gathering critical information about the use case or task environment or the AI system and determining the level of directly perceivable robustness risks associated with the AI system. This risk profile forms the foundation for subsequent risk assessments.

Understanding the level of robustness risk that the AI system poses while taking into account its scope, nature, context, and purpose is the main goal during the risk profiling stage. For generative AI use cases, this could include chatbots as virtual assistants that help with things like translating languages, writing code, making news automatically, summarising text, sending emails automatically, analysing sentiment, building knowledge bases, making synthetic datasets, analysing documents automatically, and making reports automatically.

To evaluate the robustness risk profile effectively, factors such as the scope (defining the boundaries of the AI system), nature (identifying the type of task and its relevance), context (considering the business and deployment environment), and purpose (understanding the intended use and the specific problem the AI system aims to solve) are taken into account.

By conducting a comprehensive robustness risk profile evaluation, the assessment framework assists in measured risk assessment and risk weighting approaches at the subsequent stages (drivers layer and influencer layer).

Key questions to consider for Robustness risk profile are:

1. **Scope**: What is the scope of the AI system? What are the boundaries and limitations of its functionality?
2. **Nature**: What is the nature of the task that the AI system performs? How critical or sensitive is the task in terms of potential impact?
3. **Context**: What is the context in which the AI system operates? What are the specific business and deployment environments in which it is deployed?
4. **Purpose**: What is the purpose of the AI system? What problem is it intended to solve? Are there potential for deviation or drift of the model from the intended purpose?
5. **Regulatory attention or public scrutiny**: Whether the industry or used case is known to have significant regulatory attention?

6. **Stakeholders**: Who are the intended users, beneficiaries and key stakeholders of the AI system? What is the level of competency of users who interact with the system?
7. **Robustness risk**: Whether the use case or industry environment is prone to robustness issues or failures? Provide a brief regarding exposures with reference to each of the elements of Robustness including Privacy, Security, Safety, Reliability, Resilience and Causality.
8. **Known vulnerabilities**: Does the use case contain known vulnerabilities or weaknesses that could affect the robustness of the AI system in its intended use? Provide a brief regarding potential exposures with reference to each of the elements of Robustness including Privacy, Security, Safety, Reliability, Resilience and Causality.
9. **Consequences**: What are the potential consequences or impacts of a robustness failure in the AI system or the use case? How severe could these consequences be, and who could be affected?
10. **Human agency and control**: What is the extent of human-in-the-loop or on-the-loop in the process? How critical is human oversight in this project? What is the level of human agency and control on the AI system?

These questions help identify key factors and considerations that help determine the robustness risk profile of an AI system in a specific use case context. The auditee shall respond to each of the questions briefly and provide a score for the risk profile. The score is provided using a likert scale of 0 to 5, wherein '0' represents no risk and '5' represents very high risk for each of the questions above. The overall score of the Robustness Risk profile of the use case shall be compiled (the cumulative score will be between 0 to 50). This score demonstrates the level of perceived inherent Robustness Risk associated with the AI system and shall be treated as follows:

- If the overall score is 0, then the Robustness Risk for the AI system is classified as 'Z', representing no risk.
- If the overall score is between 1 to 10, then the Robustness Risk for the AI system is classified as 'C' representing low risk.
- If the overall score is between 11 to 35, then the Robustness Risk for AI system is classified as 'B' representing moderate risk.
- If the overall score is between 36 to 50, then the Robustness risk for AI systems is classified as 'A' representing high risk.

Based on the above, the auditee shall determine if a provide risk weightage to the risks assessed in 7.2. The weightage shall be a score of 3 for high risk, 2 for medium risk and 1 for low risk. Such a weightage shall be added to each of the risks at the time of risk assessment and ranking.

In addition, the auditee shall provide a sub score for each of the elements (Privacy, Security, Safety, Reliability, Resilience and Causality) regarding questions 7 & 8. The scoring shall be a likert score of 0 to 3 wherein '0' represents no risk and '3' represents high risk. The extent of testing evaluation verification and validation shall be proportionately determined for elements that are considered high risk (increased number of tests for high risk). The rubric for determining the extent of robustness assessment is provided below:

Robustness Risk Profile	**Drivers**			**Influencers**		
	Assess	**Rank**	**Test**	**Governance**	**Report**	**Remediate**
No Risk	No	No	No	No	No	No

Low Risk	Yes	Yes	No	No	No	No
Medium Risk	Yes	Yes	Yes*	No	No	No
High Risk	Yes	Yes	Yes**	Yes	Yes	Yes*

- For Critical risks that are assessed and ranked.
- For Critical and Moderate risks that are assessed and ranked.

Refer the subsequent sections to understand the actions that need to be performed as part of drivers and influencers.

The auditee prepares a Robustness Risk profile of the use case based on the above factors and documents it as a report. The auditee considers the outcome of Robustness Risk profile of use case for further Robustness Assessment based on the above rubric. The rubric can be applied irrespective of size of organisation, and size or type of model. The robustness risk profile determines the impact of the model.

Robustness risk - Driver

The Robustness Risk Driver layer of the framework focuses on the risk assessment, ranking, and testing processes. This layer aims to address specific risks associated with the use case or AI system by conducting a comprehensive assessment. By taking a risk-based approach and utilising insights from the risk profile of the AI system, organisations can tailor their efforts in the robustness assessment.

The driver layer consists of three components, collectively referred to as ART:

1. **Assess:** This component covers approaches to assessing the risks associated with the robustness of AI systems in the context of the specific use case. It involves identifying potential risks, evaluating their likelihood and impact, and gaining an understanding of the vulnerabilities and potential threats to the AI system's robustness.

2. **Rank:** The rank component focuses on approaches for ranking and scoring the identified risks based on factors such as severity, likelihood, detectability, and relevance. By assigning priority levels to risks, organisations can determine which risks require immediate attention and allocate resources accordingly.

3. **Test:** The test component involves testing approaches for assessing the robustness of the AI system. This includes defining testing processes, metrics, and methodologies to evaluate the system's resilience, performance, and vulnerability to potential threats. It aims to identify weaknesses, potential failures, and areas for improvement in the AI system's robustness.

This approach ensures a systematic and thorough assessment of risks, enables effective risk prioritisation, and facilitates robust testing to enhance the overall robustness of the AI system.

7.1.2.1. Assess

7.1.2.1.1. Robustness Risk inputs and indicators compilation

Robustness risk inputs and indicators can be compiled in a risk log from a number of sources, including the following prominent sources:

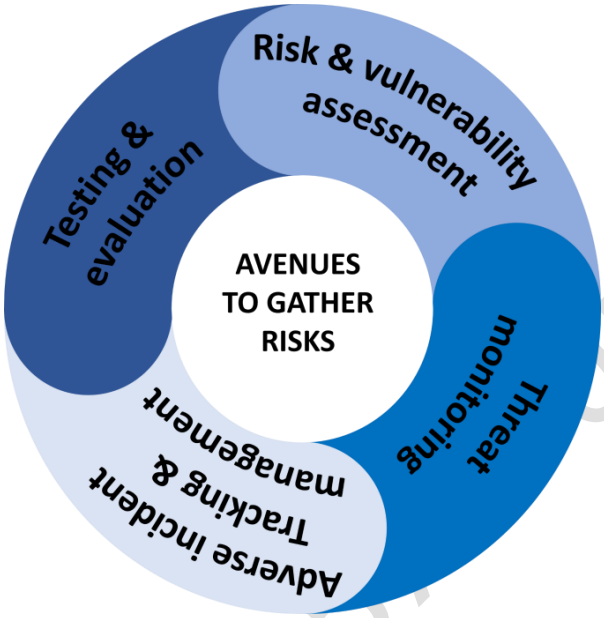

Figure 4:Robustness risk input and indicators

Risk and vulnerability assessment: risk assessment or vulnerability assessment for the AI system or similar AI systems could provide a preliminary list of potential risks to be considered in the risk log.

Threat monitoring is the process of threat modelling, scanning, and monitoring to periodically examine the perimeter for potential threats associated with the model. The risks identified through the process could be compiled in a risk log on an ongoing basis.

Adverse incident tracking mechanism provides opportunity to collect feedback from employees, customers, partners, civil society stakeholders on potential failure modes of the model and associated robustness risks. These could be compiled as part of the risk log.

Testing and evaluation is one of the important sources to suggest the potential failure, error, inconsistency in the model outputs or performance. The risks identified from the testing and evaluation process could be compiled into a risk log.

7.1.2.1.2. Robustness Risk analysis

Robustness risk analysis plays a vital role in robustness risk assessment for AI systems. It helps identify, assess, and understand potential risks, enabling organisations to develop appropriate strategies and measures to enhance the system's privacy, security, safety, reliability, resilience, and causality while maintaining trust and mitigating potential harm.

Risks associated with robustness are compiled from the avenues referred above. These risk indicators and inputs are essentially representative of the risk contributors discussed in section 5.3.

The approach starts with identifying risks by considering potential sources of vulnerabilities and threats specific to the AI system and its use case. This may involve analysing data sources, model architecture, potential attack vectors, and external factors that could impact the system's robustness.

Risk Categories The risks shall be organised under five broad categories of risk:. Robustness alignment, training and communication, 2. Robustness measures and controls, 3. Monitoring and detection, 4. Error handling and incident management, and 5. Documentation and governance.

Once risks are identified, they are assessed by evaluating their likelihood of occurrence and potential impact on the system. This assessment helps prioritise risks and determine the level of attention and resources required for mitigation. Factors such as severity, likelihood, detectability, and relevance are considered in the risk assessment process.

After assessing risks, the next step is to analyse and understand their potential consequences and implications. This involves evaluating the potential outcomes, impacts on user experience, system performance, alignment with regulation/ standard / industry guidelines, and overall trustworthiness of the AI system.

7.1.2.2. Rank

Risk ranking involves prioritising risks based on their significance and potential impact. It includes efforts to to rank and score risks based on potential implications. A brief guidance on parametric considerations for ranking the risks are provided below.

7.1.2.2.1. Ranking Risks

Relevance
1. How closely does the risk impact align with the robustness goals and objectives of the AI system? 2. To what extent does the risk impact affect the ability of the AI system to meet the robustness needs and requirements of users or stakeholders? 3. How significantly does the risk impact the overall effectiveness and reliability of the AI system in performing its intended tasks?

High Risk	Medium Risk	Low Risk
1. The risk impact significantly hinders the robustness goals and objectives of the AI system, potentially leading to critical failures or vulnerabilities. 2. The risk impact severely affects the ability of the AI system to meet the needs and requirements of users or stakeholders, resulting in substantial limitations or dissatisfaction. 3. The risk impact poses a significant threat to the overall effectiveness and reliability of the AI system, potentially compromising its performance and trustworthiness.	1. The risk impact partially hampers the robustness goals and objectives of the AI system, with some areas of improvement or potential vulnerabilities. 2. The risk impact moderately affects the ability of the AI system to meet the needs and requirements of users or stakeholders, resulting in certain limitations or suboptimal performance. 3. The risk impact poses a moderate threat to the overall effectiveness and reliability of the AI system, requiring attention and mitigation to ensure satisfactory functioning.	1. The risk impact minimally affects the robustness goals and objectives of the AI system, with limited or negligible vulnerabilities. 2. The risk impact has minimal impact on the ability of the AI system to meet the needs and requirements of users or stakeholders, resulting in satisfactory performance and outcomes. 3. The risk impact poses a low threat to the overall effectiveness and reliability of the AI system, with little or no compromise to its functionality and trustworthiness.
Severity		
1. To what extent will the results affect the user's life? 2. To what extent will the outcomes affect users' rights and freedom as per the constitutional and ethical considerations? 3. To what extent it intends to uphold the principles of availability, confidentiality, integrity, explainability, generalizability, and transparency?		
High Risk	**Medium Risk**	**Low Risk**
1. AI results may have a bearing on the safety and security of individuals or may determine their critical life decisions (e.g., disease diagnosis algorithms and autonomous vehicles). 2. AI is anticipated to impact individuals' eligibility for certain benefits, thereby significantly	1. AI results may affect individuals' convenience or financial choices. 2. AI is anticipated to impact individuals' eligibility for certain benefits to a limited extent due to public interest	1. AI results may have limited or no bearing on individuals and may not impact them physically or financially. 2. AI intends to have limited or no impact on potential access to services or opportunities (social or technical accessibility).

impacting rights or freedom. 3. AI aims to address existing societal biases.	requirements. 3. AI intends to support determining prioritisation for essential services.	

Likelihood		
1. To what degree is the occurrence expected considering historical evidence, current trends, and future projections? 2. What factors contribute to the possibility of the risk manifestation? 3. Are there preventive measures already implemented to minimise the chance of realisation?		
High Risk	**Medium Risk**	**Low Risk**
1. Historical patterns show consistent occurrences or recent events indicate imminent materialisation. 2. Preventive safeguards seem insufficient or absent altogether. 3. Multiple contributing elements align, pointing towards a heightened probability of risk emergence.	1. Occasional instances were documented previously but do not suggest regularity. 2. Mitigation strategies might reduce the chances yet cannot eliminate them entirely. 3. Few contributing aspects favour risk occurrence amidst countervailing forces.	1. Infrequent past episodes make recurrence improbable. 2. Robust protective actions diminish likelihood considerably. 3. Minimal influencing variables lean against risk appearance
Detectability		
1. At what stage can the risk be recognized before causing damage? 2. Is early warning detection available for timely intervention? 3. Does the technology provide continuous monitoring capabilities addressing evolving risks?		
High Risk	**Medium Risk**	**Low Risk**
1. Risks are noticeable at nascent stages, allowing ample time for remediation. 2. Early alert systems enable swift responses, preventing severe consequences.	1. Detection happens after minor harm has occurred; however, major damages remain avoidable. 2. Limited warning signs emerge occasionally,	1. Identification takes place only upon significant repercussions unfolding. 2. Scarcely visible indicators require expert scrutiny to discern underlying hazards.
3. Ongoing surveillance tools continuously assess shifting threats.	necessitating careful observation. 3. Periodic evaluations identify emerging risks demanding attention	3. Technology lacks comprehensive tracking capabilities, leading to unnoticed

7.1.2.2.2. Scoring Risks

The auditee could adopt any approach to score the ranked risks based on their enterprise practice or the mechanism suggested below. Mechanism (suggested) is to score the risks (determined by Relevance, Severity, Likelihood and Detectability) as follows:

Aspect	High Risk	Medium Risk	Low Risk
Relevance	3	2	1
Severity	3	2	1
Likelihood	3	2	1
Detectability	1	2	3

The overall risk score (Least 1 and Highest 81) of any risk shall be a product of scores for relevance, severity, likelihood, and detectability. The overall risk score shall be a basis of tagging risks for subsequent testing process:

Overall Risk Score	Risk tagging
1-30	Limited Risk
31-60	Moderate Risk
61-81	Critical Risk

In cases where the auditee adopts their enterprise practices for scoring of risks, they may need to clearly document the approaches adopted. The auditor must review the auditee's scoring methodology documentation along with the overall robustness assessment report.

7.1.2.2.3. Scaling Risks

The risk scores shall be scaled to 100 for each of the Risk categories described in the previous section and at AI robustness risk level to 100.

7.1.2.3. **Test**

Risks assessed and ranked are tested using various methods for robustness as part of the assessment process. The methods of robustness testing include process, metrics evaluation, threat evaluation or vulnerability testing, adversarial testing, and counterfactual testing. Such an approach ensures that the AI system's development and deployment process is robust, metrics are effectively evaluated, vulnerabilities are identified, adversarial robustness is tested, and the system's response to counterfactual inputs is assessed. The methods shall take into account distribution, outliers, boundary, or edge cases for developing test cases based on appropriateness.

The risks assessed and ranked shall be mapped with appropriate testing methods, thereby ensuring that there is a testing method for each of the risks ranked. In some cases, there may be a need to have more than one approach to testing a risk. Tests should be conducted based on priorities determined in the ranking using overall risk scores and risk tagging at individual risk levels.

The testing shall reveal that (a) controls or process measures are efficient and effective, or (b) there is a lack of or inadequate controls or process measures and/or ineffective or inefficient control or process measures. If the controls or process measures mitigate the risks (and are efficient and effective), the risks shall become zero, otherwise, the risks shall remain the same.

7.1.2.3.1. Process

Process testing involves examining the processes and controls within the system's lifecycle to identify and prevent robustness risks. It focuses on ensuring that robustness is built into the development, deployment, and maintenance processes. Some of the aspects covered in process testing include:

Robustness alignment, training and communication

Area	Aspects to cover or consider
Robustness alignment, training and communication	• Robustness strategy alignment • Personnel Training on robustness strategy • Communication and Collaboration with different stakeholders • Resource Allocation for robustness management • Redundancy Mechanism for robustness • Resource Scalability for robustness • Communication Channels for Response Coordination regarding robustness testing
Robustness measures and controls	• Data collection and quality process • Access control and user rights management

	• Robustness requirement definition • Encryption mechanism • Data protection measures • Tools and equipment calibration mechanism • Cloud security mechanism • Threat and incident tracking mechanism • Backup and recovery procedures
Monitoring and detection	• Model enhancement feedback mechanism • Model drift monitoring mechanism • Resource usage monitoring mechanism • Version management control mechanism • Logging and monitoring mechanism • Patch management process • Security incident monitoring process • Regular security drills and audits
Error handling and incident management	• Error handling mechanism • Defect, failure, or incident management mechanism • Triggers, threats, errors, and omissions monitoring mechanism • Failed data validation or transformation tracking mechanism • Failure mode feedback mechanism • Incident response plan and updates • Incident reporting mechanism
Documentation and governance	• Standardised procedures documentation • Model deployment documentation • Metrics, methods, and thresholds documentation • Metrics and testing results interpretation documentation • Post-Incident analysis documentation

7.1.2.3.2. Metric

Metrics evaluation involves assessing the robustness of the AI system based on specific metrics. The details of these metrics can be found in Section 6 of the assessment framework. This evaluation helps measure the system's performance and robustness against predetermined criteria.

Area	Aspects to cover/ consider
Robustness measures and controls	• Input data metrics covering missing data, data schema checks, outlier detection, etc • Metrics tracking corrupted inputs and pipeline bugs • Trends in data that do not meet predefined quality standards.

	• Trends in data collection sensors, tools, or equipment malfunctions or calibrations affect the quality of data. • Trends in data that do not pass validation criteria. • Trends of identified biases in data collection processes. • Trends of results that fail validation checks.
Monitoring and detection	• Trends in system failures or errors due to inadequate fault tolerance. • Trends of performance degradation during integration with other systems. • Trends of missing or incomplete logs or metrics for performance and/or quality. • Number/ proportion of missing or outdated robustness measures or metrics • Trends of insufficient or non-diverse test data. • Frequency of architectural reviews for robustness. • Trends in system functionality or scenarios not covered by testing. • Trends in testing of tools, integrations, or code for robustness.
Error handling and incident management	• Trends of incidents without clearly defined escalation procedures. • Trends in security incidents were not detected due to insufficient tools. • Trends in data security breaches or incidents. • Trends in system components or activities are not adequately monitored.

It is necessary for the auditee to determine the metrics, measures, and thresholds associated with robustness and document them, along with sufficient reasoning on why such approaches were considered appropriate for the AI system.

7.1.2.3.3. Threat/ Vulnerability

Threat Evaluation or Vulnerability Testing involves various approaches, including code reviews, unit and systems testing, configuration management reviews, and penetration tests. The objective is to identify system-level vulnerabilities associated with the AI system.

Area	Aspects to cover/ consider
Robustness measures and controls	• Penetration testing • Data encryption testing • Threat modelling • Social engineering test • Network security testing • Security patch management testing

	• Static Application Security Testing (SAST), Dynamic Application Security Testing (DAST) and Interactive Application Security Testing (IAST)

7.1.2.3.4. Adversarial

Adversarial testing involves testing the AI model for adversarial robustness. Adversarial testing involves intentionally crafting inputs that aim to deceive or manipulate the model. By subjecting the model to adversarial examples, organisations can evaluate its ability to withstand potential attacks and improve its robustness against adversarial inputs [35],[36],[37],[38],[39].

Area	Aspects to cover/ consider
Robustness measures and controls	• Optimization based attacks • Universal evasion attacks • Score based attacks • Decision based attacks • Availability poisoning • Target poisoning • Backdoor poisoning • Data reconstruction attack • Sponge attack • Membership inference • Model extraction • Property interference

7.1.2.3.5. Counterfactual

Counterfactual testing involves evaluating the AI system and its data from a robustness perspective using counterfactual inputs. Counterfactual testing explores alternative scenarios by providing modified inputs while keeping the desired outcome constant. This helps understand the system's sensitivity to changes and assess its robustness in handling different inputs.

Area	Aspects to cover/ consider
Robustness measures and controls	• Stress testing • Input validation • Robustness to changes • Resilience testing

	• Compatibility testing (including compatibility to several libraries and tools) • Performance testing assessing the impact of response times, resource utilisation, traffic / load handling etc
Monitoring and detection	• Effects of data perturbations • Model drift detection • Bias and fairness evaluation • Thresholds and alerts analysis
Error handling and incident management	• Scenario simulation • Testing error handling and exception management • Incident escalation and severity assessment • Post-incident analysis

7.1.3. Robustness risks - Influencers

The influencer layer consists of three components, namely, Governance, Reporting and Remediation. These layers help assess the potential impact of robustness. Essentially the influencer layer helps assess the implications of robustness risks in light of governance, reporting or remediation efforts within an organisation environment.

7.1.3.1. Governance

The Governance considerations shall include the following at enterprise level, development and deployment level, and use and maintenance level:

7.1.3.1.1. Enterprise

1. **Policy Development:** Does the organisation have policies and guidelines regarding the robustness of AI systems, including considerations for data quality, model development, testing, and validation?
2. **Resource Allocation:** Does the organisation allocate necessary resources, including funding, expertise, and infrastructure, to support the development and maintenance of a robust AI system?
3. **Compliance and Regulatory Alignment:** Does the organisation have processes enabling compliance with relevant laws, regulations, and ethical standards related to AI system robustness, such as data protection and privacy regulations?
4. **Vendor and Third-Party Management:** Does the organisation support implementing processes to evaluate and manage the robustness of AI systems developed by vendors or third-party providers, including due diligence in selecting trustworthy and reliable partners? Does this due diligence include AI systems?

5. **Training and Awareness:** Does the organisation provide training and awareness programs to employees and stakeholders about the importance of robustness in AI systems, including best practices, ethical considerations, and potential risks?

7.1.3.1.2. Development and deployment

1. **Data Quality:** Does the organisation have mechanisms to assess data quality, representativeness, and biases that may impact the robustness of the system?
2. **Robust Model Development:** Does the organisation have robust model development practices for developing AI models, including appropriate feature selection, preprocessing techniques, model architecture, and regularisation methods to enhance robustness?
3. **Testing and Validation:** Does the organisation have mechanisms for conducting rigorous testing and validation processes, including stress testing, edge case testing, and adversarial testing, to evaluate the robustness and performance of the AI system?
4. **Documentation and Version Control:** Does the organisation have a mechanism to maintain proper documentation and version control (including user documentation) of the AI system development process, allowing for traceability and reproducibility of results, and facilitating robustness enhancements and audits?

7.1.3.1.3. Use and maintenance

1. **Responsible Use:** Does the organisation have mechanisms to support responsible and ethical use of AI systems, including understanding the limitations, potential biases, and risks associated with the system's robustness?
2. **Feedback and Reporting:** Does the organisation have a mechanism for providing feedback on AI system performance and reporting any issues, adverse incidents, or concerns related to the robustness of the system to the appropriate teams for further investigation and action?
3. **Continuous Monitoring:** Does the organisation have a mechanism to enable active performance monitoring and robustness of the AI system during its usage, promptly reporting any anomalies or unexpected behaviours that may indicate a lack of robustness?
4. **User Training and Support:** Does the organisation have sufficient mechanisms for providing adequate training and support to users to help them understand the AI system's robustness and its limitations, enabling them to make informed decisions and utilise the system effectively?
5. **Equipping human-in-the-loop:** Does the organisation have sufficient mechanisms to train and equip human-in-the-loop or persons with human oversight to be able to prevent or limit potential catastrophic failures of the AI system?

7.1.3.1.4. Reporting

1. **Disclosure mechanism:** Does the organisation have sufficient disclosure mechanisms, including model cards and data cards, to express the limitations or failure modes of the AI systems for the users?
2. **Active stakeholder engagement process:** Does the organization actively engage with stakeholders to understand their perspectives, needs, and concerns related to AI systems? Does it include mechanism or process to for soliciting feedback and incorporating stakeholder input into decision-making and system development?
3. **Board reporting process:** Does the organization ensure that the board reporting process provides sufficient transparency and oversight of AI systems including information or performance or risk metrics and ethical considerations associated with AI systems?

7.1.3.1.5. Remediation

1. **User friendly interfaces and clear instructions:** Does the organization ensure that its user interfaces are designed with simplicity, intuitiveness, and accessibility in mind to provide a user-friendly experience? Does the organization take into account user feedback and usability testing to continuously improve the clarity and effectiveness of its instructions and interface design?
2. **Transparent disclosure on data collection and usage:** Does the organization have mechanisms to ensure that users are fully informed about the types of data collected, the purposes for which it is collected, and any potential third-party sharing or transfer of their data? Does the organization provide the user with appropriate channel for exercising their data subject rights?
3. **Responsive and accessible customer support channels:** Does the organization have customer support channels that are readily available and accessible to users, including those with different abilities or language preferences? Does the organization consistently measure the effectiveness of the customer experience in these support channels?

8.0 Mitigation Framework for Robustness Risks

Mitigating robustness risks in AI systems requires a multi-pronged approach that combines various techniques and strategies. The following mitigation strategies should be considered and implemented as appropriate:

8.1. Robust Training

Robust training techniques aim to improve the resilience of AI models by exposing them to a diverse range of perturbations, adversarial examples, or distributional shifts during the training process. These techniques can greatly enhance the robustness of the trained models to various types of perturbations and distributional shifts encountered during deployment.

8.1.1. Adversarial Training

Adversarial training involves incorporating adversarial examples, which are carefully crafted input perturbations designed to fool the model, into the training data. By training on these adversarial examples, the model learns to be more robust against adversarial attacks. Popular adversarial training methods include Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Adversarial Weight Perturbation [40], [41].

8.1.2. Data Augmentation

Data augmentation is a technique that artificially increases the diversity of the training data by applying various transformations, such as rotations, flips, noise injections, or style transfers. This exposure to a broader range of data during training can improve the model's robustness to input perturbations and distributional shifts [42], [43].

8.1.3. Domain Randomization

Domain randomization is particularly useful for robustness in simulated environments or robotics applications. It involves randomizing the simulation parameters, such as lighting conditions, textures, or object positions, during training. This exposure to diverse simulated environments can improve the model's ability to generalize to real-world scenarios [44], [45].

8.1.4. Distributionally Robust Optimization

Distributionally robust optimization (DRO) is a training paradigm that aims to optimize the model's performance across a range of potential data distributions, rather than just the empirical training distribution. This can improve the model's robustness to distributional shifts encountered during deployment [46], [47].

8.2. Model Architecture for Robustness

Certain model architectures and design principles can inherently promote robustness by incorporating structural properties or inductive biases that enhance resilience to perturbations or distributional shifts.

8.2.1. Ensemble Methods

Ensemble methods combine multiple models, such as neural networks or decision trees, to improve overall robustness. By aggregating the predictions of diverse models, ensemble methods can mitigate the individual weaknesses of each model and provide more robust predictions [48],[49].

8.2.2. Bayesian Neural Networks

Bayesian neural networks (BNNs) model the uncertainty in the network's weights by representing them as probability distributions rather than point estimates. This explicit representation of uncertainty can improve the model's robustness to distributional shifts and out-of-distribution inputs [50], [51].

8.2.3. Robust Feature Representations

Developing robust feature representations that are invariant to certain types of perturbations or distributional shifts can enhance the overall robustness of the AI system. For example, in computer vision tasks, features that are invariant to rotations, translations, or lighting conditions can improve robustness [52], [53].

8.2.4. Architectures with Built-in Invariances

Certain model architectures, such as convolutional neural networks (CNNs), inherently possess built-in invariances that can promote robustness. For instance, CNNs exhibit translation invariance, which can improve robustness to spatial perturbations in image data [54], [55].

8.3. Monitoring and Adaptation

Continuous monitoring and adaptation are essential for maintaining robustness in dynamic environments where the data distribution or operational conditions may change over time.

8.3.1. Online Monitoring

Implementing online monitoring systems that continuously track the performance and robustness metrics of the deployed AI system can help detect potential degradations or failures. This monitoring can trigger alerts or mitigation actions when robustness issues are detected [56], [57].

8.3.2. Continuous Retraining and Adaptation

As new data becomes available or distributional shifts are detected, continuous retraining and adaptation of the AI model can help maintain its robustness. This can involve periodically updating the model with new data or fine-tuning it with the latest data distribution [58], [59].

8.3.3. Anomaly and Out-of-Distribution Detection

Incorporating anomaly detection and out-of-distribution detection mechanisms into the AI system can help identify inputs or scenarios that deviate significantly from the training distribution. These detections can trigger appropriate mitigation actions, such as requesting human intervention or falling back to a more conservative mode of operation[60],[61].

8.4. Human - AI Collaboration

Involving human experts in the loop can enhance the robustness of AI systems by leveraging human intelligence, domain knowledge, and cognitive abilities that complement the strengths of AI models.

8.4.1. Human-in-the-Loop Decision-Making

In critical applications or high-risk scenarios, incorporating human judgment and decision-making in the loop can mitigate potential robustness issues by allowing human experts to override or correct the AI system's outputs when necessary[62],[63].

9.0 Rating Methodology

As can be seen in Chapter 7, the standard consists of three key measures, namely, Use case AI Robustness Risk Profile, Results of Assess-Rank-Test of Robustness Risk, and compiled insights on Govern-Report-Remediate. Of the above 3, only the first two relate to the AI system in specific and the last one is at an organisational level. Hence, the last one will be used as an indicator and not be used for rating as the standard is for rating of Robustness of AI systems.

Robustness risk profile has four measurements represented by Z (No risk), A (high Risk), B (Medium Risk) and C (low risk). And, Results of Assess-Rank-Test is a numeral value between 0-100.

The robustness risk profile and the results of Assess-Rank-Test is used for formulating the following rating:

Rating methodology		Overall Risk Score results from Access-Rank-Test		
		0-40	41-80	81-100
Robustness Risk Profile	Z (No)	Z++	Z+	Z
	A (High)	A++	A+	A
	B (Medium)	B++	B+	B
	C (Low)	C++	C+	C

'+' Denotes that the AI system risks are better managed

This approach enables context weighted measures of assessing the robustness and comparing it with other peer groups across different contextual scenarios.

In addition to the above rating, the auditee shall provide specific score for compiled insights on Govern-Report-Remediate as discussed in Chapter 7.

10.0 Abbreviations

Abbreviation	Expansion
AI	Artificial Intelligence
BNN	Bayesian Neural Network
CNN	Convolutional Neural Network
DAST	Dynamic Application Security Testing
DRO	Distributionally Robust Optimization
FGSM	Fast Gradient Sign Method
GenAI	Generative Artificial Intelligence
IAST	Interactive Application Security Testing
LLM	Large Language Model
LIME	Local Interpretable Model-Agnostic
PGD	Projected Gradient Descent
SAST	Static Application Security Testing
SHAP	Shapley Additive Explanations
SOP	Standard Operating Procedure

11.0 Acknowledgements

- Mr. Avinash Agarwal, DDG, TEC, Department of Telecommunications - Chair
- Dr. Dhawal Gupta, Group Business Director - Public Policy, Chase India
- Dr. Gopalakrishnan, Principal, School of Artificial Intelligence, Bengaluru, Amrita Vishva Vidyapeetham
- Ms. Kavita Bhatia, Scientist-G, AI & Emerging Technologies Group, MeitY
- Mr. Krishnan Narayanan, Research Lead at CeRAI, IIT Madras and Co-founder Itihaasa Research and Digital
- Mr. M Saravanan, Principal Researcher, Ericsson
- Mr. Manoj K. Parmar, CEO, CTO, AIShield, Bosch
- Ms. Sarvjeet Kaur, Scientist-G, SAG, DRDO
- Dr. Shiv Kumar, Principal Advisor, BIF
- Dr. Sunayana Sitaram, Principal Researcher, Microsoft Research India
- Mr. Sundar Narayanan, Researcher, AI & Tech Ethics
- Dr. Vijay Arya, Senior Researcher, IBM Research

12.0 References

[1]. ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE. ISO. (2014). ISO/IEC 25000:2014 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE

[2]. ISO/IEC 25059:2023 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems.ISO.(2023). https://www.iso.org/standard/80655.html

[3]. ISO/IEC 25010:2011(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.ISO.(2011).ISO/IEC 25010:2011 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models

[4]. ISO/IEC 25010:2023(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model.ISO.(2023).ISO/IEC 25010:2023 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model

[5]. X.1631 : Information technology - Security techniques - Code of practice for information security controls based on ISO/IEC 27002 for cloud services.ITU.(2015).X.1631 : Information technology - Security techniques - Code of practice for information security controls based on ISO/IEC 27002 for cloud services (itu.int)

[6]. X.800 : Security architecture for Open Systems Interconnection for CCITT applications.ITU.(1991).X.800 : Security architecture for Open Systems Interconnection for CCITT applications (itu.int)

[7]. ISO/IEC 2382:2015 Information technology — Vocabulary.ITU.(2015).ISO/IEC 2382:2015 - Information technology — Vocabulary

[8]. ISO 16484-5:2022 Building automation and control systems (BACS).ITU.(2022).ISO 16484-5:2022 - Building automation and control systems (BACS) — Part 5: Data communication protocol

[9]. ETSI GS ZSM 012 V1.1.1 Zero-touch network and Service Management(ZSM); Enablers for Artificial Intelligence-based Network and Service Automation.ETSI.(2022) ETSI - ZSM - Zero touch network & Service Management

[10]. ISO/TR 25060:2023 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — General framework for Common Industry Format (CIF) for usability-related information.ISO.(2023).ISO/TR 25060:2023 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — General framework for Common Industry Format (CIF) for usability-related information

[11]. ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.ISO.(2022).ISO/IEC 22989:2022 - Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

[12]. ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements.ISO.(2019).ISO 22301:2019 - Security and resilience — Business continuity management systems — Requirements

[13]. ISO/IEC Guide 51:2014 Safety aspects — Guidelines for their inclusion in standards. ISO.(2014).ISO/IEC Guide 51:2014 - Safety aspects — Guidelines for their inclusion in standards

[14]. ISO/IEC 27000:2018 Information technology — Security techniques — Information security management systems — Overview and vocabulary. ISO.(2018).ISO/IEC 27000:2018 - Information technology — Security techniques — Information security management systems — Overview and vocabulary

[15]. M.3016.0 : Security for the management plane: Overview.ITU.(2005).M.3016.0 : Security for the management plane: Overview (itu.int)

[16]. NIST-SP-800-30 Guide for Conducting Risk Assessments. NIST. (2012).nistspecialpublication800-30r1.pdf

[17]. X.1361 : Security framework for the Internet of things based on the gateway model.ITU.(2018).X.1361 : Security framework for the Internet of things based on the gateway model (itu.int)

[18]. CCI. (2021). MARKET STUDY ON THE TELECOM SECTOR IN INDIA Key Findings and Observations. market-study-on-the-telecom-sector-in-india1652267616.pdf (cci.gov.in)

[19]. Purushothaman KG. (2023). Indian telecom industry in 2023: Setting on the pathway to global success, ET Telecom (indiatimes.com)

[20]. ETTelecom2023 - Muntazir Abbas. (2023). Telecom Diary: Rise in premium smartphone sales may augur well for India's 5G adoption, ET Telecom (indiatimes.com)

[21]. Mahmoud, H. H. H., & Ismail, T. (2020, December 1). A Review of Machine learning Use-Cases in Telecommunication Industry in the 5G Era. IEEE Xplore. https://doi.org/10.1109/ICENCO49778.2020.9357376 .

[22]. Ramiro, J., & Khalid Hamied. (2011). Self-Organizing Networks. John Wiley & Sons.

[23]. AI Use Cases in Telecom Relevant for 2022 with 8 examples. (2021, December 12). MindTitan. https://mindtitan.com/resources/industry-use-cases/artificial-intelligence-in-telecom-business/

[24]. Maatouk, A., Piovesan, N., Ayed, F., De Domenico, A., & Debbah, M. (2023). Large language models for telecom: Forthcoming impact on the industry. arXiv preprint arXiv:2308.06013.

[25]. Holm, H., Gunnarsson, M., Nimara, D. D., Wei, J., Gebre, F. G., & Huang, V. (2022, January 20). Adopting neural language models for the telecom domain. Ericsson.com. https://www.ericsson.com/en/blog/2022/1/neural-language-models-telecom-domain

[26]. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

[27]. How telcos could use gen AI to revitalize profitability and growth | McKinsey. (n.d.). Www.mckinsey.com. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/how-generative-ai-could-revitalize-profitability-for-telcos

[28]. Wu, W., Zhou, C., Li, M., Wu, H., Zhou, H., Zhang, N., Xuemin, Shen, & Zhuang, W. (2021). AI-Native Network Slicing for 6G Networks. ArXiv:2105.08576 [Cs]. https://arxiv.org/abs/2105.08576

[29]. Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety, 215, 107864. https://doi.org/10.1016/j.ress.2021.107864

[30]. Cheng, W., Luo, E., Tang, Y., Wan, L., & Wei, M. (2021). A Survey on Privacy-security in Internet of Vehicles. 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). https://doi.org/10.1109/dasc-picom-cbdcom-cyberscitech52372.2021.00109

[31]. Soori, M., Arezoo, B., & Dastres, R. (2023). Digital Twin for Smart Manufacturing, A Review. Sustainable Manufacturing and Service Economics, 2, 100017. https://doi.org/10.1016/j.smse.2023.100017

[32]. Joe, B., Park, Y., Hamm, J., Shin, I., & Lee, J. (2022). Exploiting missing value patterns for a backdoor attack on machine learning models of electronic health records: Development and validation study. JMIR Medical Informatics, 10(8), e38440.

[33]. Galloway, A., Golubeva, A., Tanay, T., Moussa, M., & Taylor, G. W. (2019). Batch normalization is a cause of adversarial vulnerability. arXiv preprint arXiv:1905.02161.

[34]. Baev, R. V., Skvortsov, L. V., Kudryashov, E. A., Buchatskiy, R. A., & Zhuykov, R. A. (2022). Preventing Vulnerabilities Caused by Optimization of Code with Undefined Behavior. Programming and Computer Software, 48(7), 445-454.

[35]. Improper Data Validation.OWASP Foundation. Improper Data Validation | OWASP Foundation

[36]. OWASP Machine Learning Security Top Ten.OWASP Foundation.(2023).OWASP Machine Learning Security Top Ten | OWASP Foundation

[37]. Sadeghi, K., Banerjee, A., & Gupta, S. K. (2020). A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE transactions on emerging topics in computational intelligence, 4(4), 450-467.

[38]. Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning. Gaithersburg, MD.

[39]. The taxonomy of security threats towards machine learning.R esearchGate.(2018).The taxonomy of security threats towards machine learning. | Download Scientific Diagram (researchgate.net)

[40]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR).

[41]. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L. S., & Goldstein, T. (2019). Universal adversarial training. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 5636-5643.

[42]. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.

[43]. Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 702-703.

[44]. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 23-30.

[45]. Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., ... & Birchfield, S. (2019). Structured domain randomization: Bridging the reality gap by context-aware synthetic data. Proceedings of the International Conference on Robotics and Automation (ICRA), 7249-7255.

[46]. Shafieezadeh-Abadeh, S., Kuhn, D., & Esfahani, P. M. (2019). Distributionally robust logistic regression. Advances in Neural Information Processing Systems (NeurIPS), 32.

[47]. Duchi, J. C., & Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. arXiv preprint arXiv:1810.08750.

[48]. Dietterich, T. G. (2000). Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, 1-15.

[49]. Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

[50]. Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning (ICML), 1050-1059.

[51]. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision?. Advances in Neural Information Processing Systems (NeurIPS), 5574-5584.

[52]. Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and regularization of support vector machines. Journal of Machine Learning Research, 10(7).

[53]. Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., & Babu, R. V. (2016). An introduction to deep convolutional neural nets for computer vision. Deep Learning for Visual Computing, 107-176.

[54]. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[55]. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. Advances in Neural Information Processing Systems (NeurIPS), 3856-3866.

[56]. Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. Advances in Neural Information Processing Systems (NeurIPS), 5541-5552.

[57]. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations (ICLR).

[58]. Kuznetsov, V., & Mohri, M. (2015). Towards non-degenerate parametric rectifiers for linear and non-linear deep neural networks. arXiv preprint arXiv:1505.03866.

[59]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. International Conference on Machine Learning (ICML), 1321-1330.

[60]. Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2019). Deep anomaly detection with outlier exposure. International Conference on Learning Representations (ICLR).

[61]. Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. International Conference on Learning Representations (ICLR).

[62]. Raghu, M., Irvin, J., Julien, C., Recht, B., & Le, Q. V. (2020). Human-in-the-loop machine learning. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), 4808-4812.

[63]. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

13.0 Annexure - I

Category	Libraries & Tools
Data Manipulation & Processing	• **NumPy** – Numerical computing with support for arrays, matrices, and mathematical functions. • **Pandas** – Data analysis and manipulation using DataFrames and Series. • **Dask** – Parallel computing for large datasets. • **Polars** – Fast DataFrame library optimized for performance. • **Vaex** – Efficiently processes large tabular datasets.
Data Visualization	• **Matplotlib** – Basic plotting library (line plots, bar charts, scatter plots, etc.). • **Seaborn** – Statistical data visualization (heatmaps, pair plots, etc.). • **Plotly** – Interactive visualizations for dashboards and web apps. • **Bokeh** – Interactive and real-time visualizations. • **ggplot (plotnine)** – Grammar of Graphics-style visualization similar to R's ggplot2.
Machine Learning	• **Scikit-learn** – Core machine learning library (classification, regression, clustering, etc.). • **XGBoost** – Optimized gradient boosting library. • **LightGBM** – Fast gradient boosting for large datasets. • **CatBoost** – Gradient boosting optimized for categorical features. • **H2O.ai** – AutoML and scalable ML models.
Deep Learning	• **TensorFlow** – Google's deep learning framework. • **PyTorch** – Deep learning framework with dynamic computation graphs. • **Keras** – High-level API for TensorFlow. • **MXNet** – Scalable deep learning framework by Apache.
Statistical Analysis & Hypothesis Testing	• **SciPy** – Scientific computing (statistical tests, optimization, signal processing). • **Statsmodels** – Statistical modeling, hypothesis testing, and regression. • **Pingouin** – Advanced statistical analysis. • **PyMC** – Bayesian statistical modeling.
Feature Engineering & Data Cleaning	• **Feature-engine** – Feature engineering methods. • **Scipy.stats** – Probability distributions and statistical functions. • **Missingno** – Visualization of missing data. • **Imbalanced-learn** – Handling imbalanced datasets. • **Category Encoders** – Encoding categorical variables.
NLP(Natural Language Processing)	• **NLTK** – Traditional NLP tasks. • **SpaCy** – Industrial-strength NLP. • **Transformers (Hugging Face)** – Pretrained transformer models. • **Gensim** – Topic modeling and word embeddings. • **TextBlob** – Simple NLP processing.

Time Series Analysis	• **Statsmodels** – Time series forecasting methods. • **Prophet (Facebook)** – Time series forecasting with trend and seasonality. • **tsfresh** – Feature extraction for time series. • **Darts** – Advanced time series forecasting.
Anomaly Detection	• **PyOD** – Outlier detection algorithms. • **ELI5** – Explainability of anomaly detection models. • **Luminol** – Anomaly detection in time series.
Data Scraping & Processing	• **BeautifulSoup** – Web scraping for HTML and XML. • **Scrapy** – Web scraping framework. • **Requests** – HTTP requests handling. • **Lxml** – XML and HTML parsing.
Big Data & Distribution Computing	• **Apache Spark (PySpark)** – Distributed computing for big data. • **Dask** – Parallel computing. • **Modin** – Faster Pandas alternative.
AutoML (Automated Machine Learning)	• **Auto-sklearn** – AutoML based on Scikit-learn. • **TPOT** – AutoML using genetic programming. • **H2O.ai** – Scalable and fast AutoML.
Explainability & Interpretability	• **SHAP** – Explain model predictions. • **LIME** – Model interpretability.

Annexure-II

Template for submitting Comments or Feedback

[Comments on each section/sub section/table/figure etc. of the draft TEC 57076:2025, be stated in a fresh row. Information/comments should include reasons for comments and suggestions for modified wordings of the clause]

Name of Commentator/Organization …….

S. No.	Section of the Draft Standard	Clause/Para/Table/ Figure No. of draft Standard	Comments/ Suggested modified Wordings	Justification for proposed Change
1.				
2.				
3.				
4.				
5.				
6				
7.				
8.				
9.				
10.				

Note- Kindly insert more rows as necessary for each clause/table, etc.

Name:
Email:
Mobile: