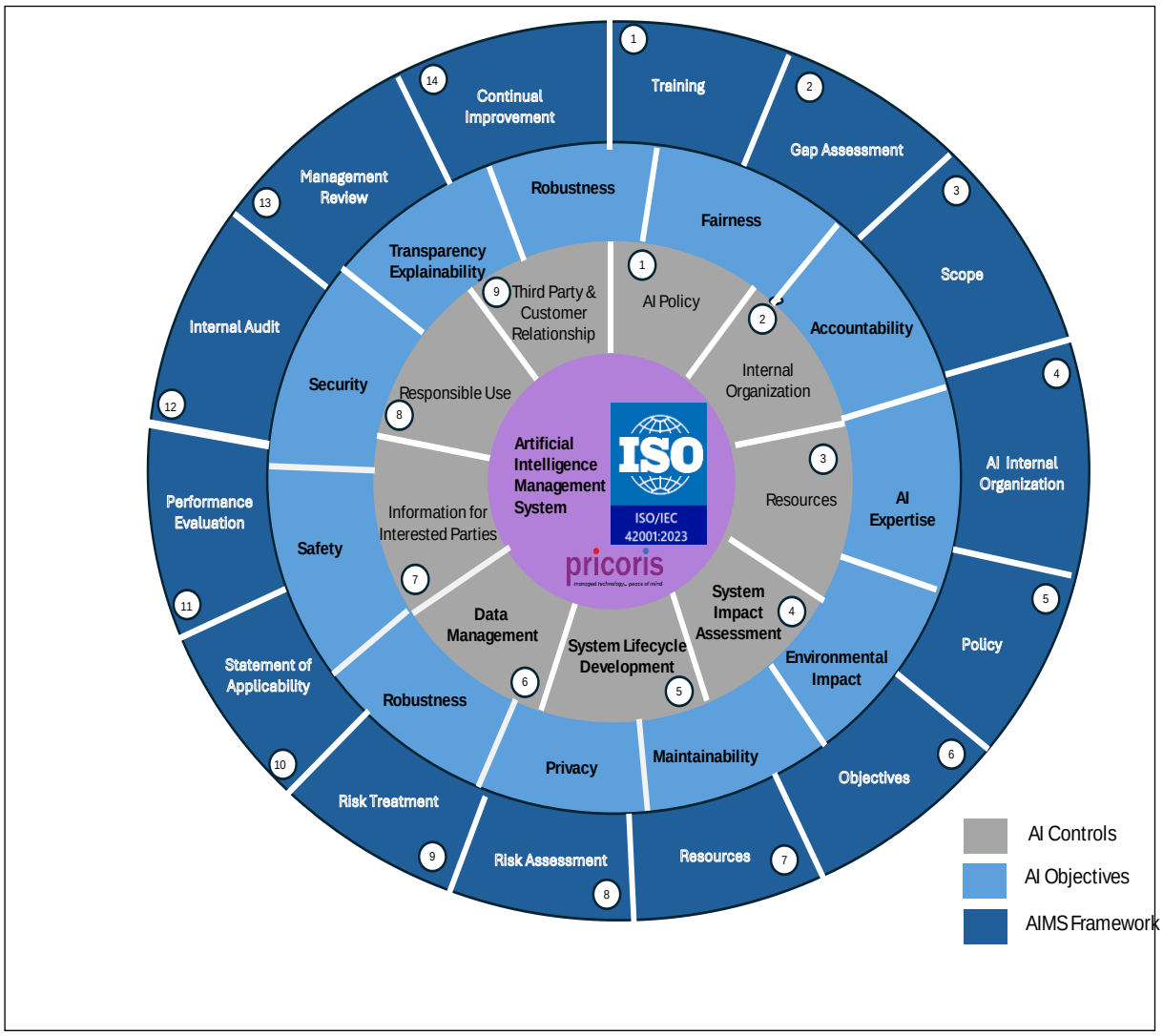
Artificial Intelligence Management System – ISO/IEC 42001:2023



Diagrammatic Representation of AIMS ISO/IEC 42001:2023

#implementationISO42001, #ISO42001, #AIMS, #ArtificialIntelligence #ResponsibleAI, #ObjectivesofAI, #LeadImplementerISO42001 #PrinciplesofAI, #trainingAIMS, #ArtificialIntelligenceManagementSystem, #RAIM

Responsible AI Objectives-AIMS

Artificial Intelligence Management System – ISO/IEC 42001:2023

AI Objectives- the middle layer of AIMS

Part 2 of 3 of Artificial Intelligence Management System

1. Are you assured that your AI algorithms are making the most effective decisions?
2. Do you have mechanisms in place to manage your AI's influence on society?
3. Are you comfortable with the potential risks involved in utilizing third-party AI systems
4. Is your organization prepared to adapt to evolving AI regulations and standards?
5. Can you effectively monitor and control the ethical implications of your AI systems?

If answer to any one of these questions is no, please consider ISO 42001 – Artificial Intelligence Management System.

- ISO 42001 is a best practise framework which sets the baseline for organizations which use, develop, monitor or provide products or services that utilize AI). – Introduction to ISO 42001:2023).
- For Part 1 of the article Click here https://www.linkedin.com/feed/update/urn:li:activity:7208141161989832704/
- **This is Part 2 of the 3 series of articles on AIMS.**

AI technologies are revolutionizing various sectors, but their deployment also raises significant ethical concerns. Addressing these concerns is crucial for ensuring AI systems are fair, transparent, and beneficial to all. Below, we explore several key ethical principles in AI, along with real-world examples that highlight the importance of each principle. These could be referred to as Responsible AI Pillars, Responsible AI Objectives, Responsible AI Principles. Some of these are mentioned in the second circle of our diagram.

Accountability

In simple terms, when AI system takes decisions, who is to be held accountable for these decisions? Uncertain accountability frameworks can cause operational hazards, legal concerns, and brand damage. These stakeholders can range from the users, their companies, AI developers, AI vendors, training data providers and regulatory bodies.

Environmental Impact

The environmental impact of AI systems encompasses various factors related to the energy consumption, resource use, and overall ecological footprint associated with the development, deployment, and operation of AI technologies. Key aspects could

include energy consumption by data centres during model training, circular economy of hardware, carbon footprint, obsolescence and e-waste.

Fairness

Fairness in AI refers to attempts to correct algorithmic bias in automated decision processes using machine learning models. If computers used sensitive variables to make machine-learning decisions, such as gender, race, sexuality, or disabilities, it may result in biased or unfair outputs. Biases can be introduced in the model in various stages –

1. **Business problem formulation** – Fairness can be compromised if the use case is not clear, limitations and exceptions are not defined, and people accountable in developing the AI-ML model are not known.

2. **Training Data** – Biases that are introduced from the training and validation data as a result of unfairness inherent in the data. Data should be high quality and relevant to the problem.
    a. **Bias in Sample data–** If the training dataset is not representative of the population. Check for issues in sampling, presence of hidden proxy variables, legal or consent issues. Examples include racial, gender or political bias.
    b. **Identifying and transforming sensitive features –** If a disfavoured group is present in the sample, then changing the model weights such that outcome for this disfavoured group is changed.
3. **Training Algorithm / model architecture** – Biases that are introduced as a result of the design of the model, giving misleading results even from good data.
    a. **Model Build** – Model architecture itself can have inherent problems, resulting in biases like Regression bias, Classification bias, Clustering bias etc. There can be calculation errors in model parameters, resulting in overfitted/underfitted models, which introduce bias and noise in output data.
    b. **Model Drift** – the business problem may evolve over time, thus AI-ML model might go stale and require re-modelling and re-training over time, re-introducing biases mentioned above at every stage.

There are mathematical models available to measure fairness, like statistical independence, confidence intervals, separation study and sufficiency study.

Privacy

AI uses vast amounts of data, raising questions regarding collection, processing, and use. As AI advances, privacy and data protection issues grow. Misuse or disclosure of sensitive data may harm subjects. Adversarial attacks and model poisoning allow malicious actors to acquire sensitive data, change AI-driven choices, or impair AI system integrity and reliability, compromising privacy and security. AI-powered facial recognition and location tracking pose mass surveillance and privacy concerns. These technologies can track people's habits, actions, and movements, compromising privacy and civil freedoms. Privacy design in AI lifecycle, data anonymization and minimization, and compliance with rules and standards can solve such problems.

Maintainability

Maintainability refers to the organization's ability to adapt the AI system to address problems or adapt to new needs.

- **Modularity and reusability** – Modular design of AI systems allow components of pipeline to be reused/replaced without affecting the entire system.
- **Documentation and version control-** Comprehensive and up-to-date documentation, including code comments, user manuals, and documentation of data preprocessing, model training, and deployment processes. Using version control systems helps in tracking changes, collaborating with others, and reverting to previous versions if needed.
- **Code quality** - High-quality code that follows best practices, including proper naming conventions, consistent style, and efficient algorithms, makes maintenance easier.
- **Scalability** - The system should be designed to handle increased loads and to scale up or down as needed, which also involves maintaining performance and resource management.

Robustness

AI Robustness refers to being resilient to input data or model parameter changes, ensuring consistent and reliable performance in uncertain or unexpected scenarios. Robustness improves reliability and resilience of the model to dynamic changes in data, but also has drawbacks of overly conservative models which can decrease fairness of the model.

Model robustness is threefold,

i) **Data pipeline**, where data validation modules are in place to check peculiarities in future data,

ii) **Model pipeline**, where we want to ensure the model cannot be attacked to produce undesirable output, and

iii) **System robustness**, since the model will be integrated to some applications downstream, the entire pipeline should be secure.

Quantitative metrics to assess the robustness of AI systems focus on evaluating how well the model performs under various stress conditions, including adversarial attacks, noise, and data distribution shifts. Key metrics are False positives/negatives rate, MSE study, Wasserstein distance and Brier score.

Safety

Safety is the expectation that a system will not risk human life, health, property, or the environment under specific situations. In applications where safety is paramount (e.g., autonomous vehicles, healthcare), the AI system should undergo rigorous testing and validation to ensure it meets safety standards and regulatory requirements. There should be multiple fail-safes and redundancies to handle failures gracefully without causing harm or significant disruption. With Human in the Loop (HITL), AI system can enhance safety, especially in critical decision-making scenarios. Humans can intervene, override, or guide the AI's actions when necessary.

Security

Security in AI systems involve protecting the AI models, data, and infrastructure from threats and vulnerabilities. This includes ensuring the confidentiality, integrity, and availability of the system and its data.

Data Security refers to ensuring data used by AI system is not tampered with, and safe from unauthorized access.

Model security refers to protecting AI model from being reverse-engineered or poisoned by unwanted parties. This includes protecting the model and data from adversarial attacks, where malicious actors manipulate inputs to deceive the model.

Techniques to mitigate these attacks include adversarial training, input validation, and anomaly detection. Other mitigation techniques include access control, audit and monitoring, risk assessment and patch management.

Transparency

AI transparency means being able to see what data AI systems use, how they make decisions, and why they do the things they do.
Transparency has two perspectives, **the organisation and the users**. Organisation perspective means the people who are in charge of making and maintaining the models and the data streams. User perspective means to know where the data comes from, what form it takes, and how the organization uses it.
There are three levels to transparency,
Algorithmic transparency focusses on the logic and model,
Interaction transparency on the user interface and
Social transparency on the impact of this interaction on society.

Explainability

As AI-ML technologies evolve and become more complicated, humans struggle to grasp and retrace AI solutions' decision-making processes. Explainability allows AI models to be evaluated to ensure that the decision-making process can be tracked in a human-friendly manner. Explainability also allows organisations to understand the models from both technical and business standpoint, thus building confidence in the AI-ML models.
Model Explainability can be measured through Tools like LIME and SHAP, which work on analysing the input-output predictions by the models.

AI Governance

AI governance involves considering the ethical implications of integrating human components into models and applications, and ensuring that the technology does not harm society or users. This involves AI developers, users and legislators to ensure that AI-related technologies are produced and used in accordance with society's values, overseeing bias, privacy, and misuse while promoting innovation and trust.

Reliability

Consistently reliable and valid AI outputs direct effective decisions and reinforce stakeholders' trust in the organization. AI systems – especially those with no or limited human input – are better equipped to minimize risks, deliver consistent outcomes, and maintain performance excellence when they are trained, validated, monitored, transparent, and undergo continuous improvement.

Accessibility

AI systems that are accessible to individuals with diverse abilities and needs. This involves designing interfaces that are user-friendly, considering factors such as language diversity and accessibility for individuals with disabilities.
User empowerment emphasizes providing users with the tools and information needed to make informed decisions about AI interactions. Developers should prioritize education and transparency, enabling users to understand how AI systems work and empowering them to control their interactions. By focusing on accessibility and user empowerment, developers can foster a positive and inclusive AI experience for all users.