ISO/IEC Standards and Design of an Artificial Intelligence System

Alessandro Simonetta[1,*,†], Maria Cristina Paoletti[2,†]

[1] Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy
[2] Professional Association of Italian Actuaries, Rome, Italy

Abstract

Artificial Intelligence (AI) is rapidly becoming a transformative force across various sectors, including healthcare, industrial automation, data processing, and finance. While AI opens new possibilities for innovation and efficiency, it also raises important concerns regarding safety, reliability, and ethical use. This article highlights the role of technical standardization in guiding responsible and transparent AI deployment, focusing on key ISO/IEC standards and introducing an ethics-by-design approach. By embedding trustworthiness into AI systems, this methodology supports compliance with standards, enhances user confidence, and encourages successful adoption of AI technologies.

Keywords

AI, Machine Learning, AI ethics, ethics by design, ISO/IEC 42001, ISO/IEC 5259, ISO/IEC 25059, ISO/IEC 25012

1. Introduction

Artificial Intelligence is quickly establishing itself as one of the most transformative technologies of our era, with applications ranging from healthcare to industrial automation, from educational to the security sector. This rapid advancement of AI brings vast opportunities for innovation and efficiency, yet also raises critical questions about the safety, reliability, and ethics of these technologies. To address these challenges, technical standardization has become a crucial tool, ensuring that AI systems are designed, developed, and deployed responsibly and transparently. This framework of standards plays an essential role in fostering the safe and ethical adoption of AI, promoting principles of quality, security, and accountability that help shape a trustworthy future for this powerful technology. The article describes the main ISO/IEC standards involved in the use of Artificial Intelligence and a new methodological approach based on ethics-by-design. Ensuring that an AI system is inherently trustworthy promotes compliance with standards and at the same time helps to increase user confidence by facilitating its adoption and success.

2. Standards ISO/IEC and AI

Over the past few years, artificial intelligence (AI) has achieved extraordinary notoriety, becoming a dominant topic in the media, surpassing any interest shown in the past.

In response to the advances and challenges posed by Artificial Intelligence, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) established a set of standards, culminating in the formation of JTC1/SC 42, which resulted in ISO/IEC 42001 [1].

This standard is designed to apply to enterprises of all sizes and industries, and its purpose is to define the criteria for creating, implementing, maintaining and continuously improving an AI management system, ensuring that AI systems are developed and used in a responsible manner. The intent of ISO/IEC 42001 is to promote the development and use of reliable, transparent and knowledgeable AI systems, emphasizing the importance of ethical principles such as fairness, non-discrimination and respect for privacy.

The integration of the ISO/IEC 25012 [2] data quality standard with the risk management process defined by ISO 31000 [3] offers a useful approach to reducing the risks associated with the use of incomplete or unbalanced datasets [4], contributing to the reduction of discrimination in artificial intelligence systems. This combination makes it possible to identify and deal with risks associated with the presence of unbalanced data in [5] training datasets.

However, it is important to remember the other ISO standards, both those already published and those in the process of being released, which can offer further guidance and inspiration for how best to structure an AI system:

IWESQ 2024: 6th International Workshop on Experience with SQuaRE family and its Future Direction
* Corresponding author.
† These authors contributed equally.
alessandro.simonetta@gmail.com (A. Simonetta);
mariacristina.paoletti@gmail.com (M. C. Paoletti)
0000-0003-2002-9815 (A. Simonetta); 0000-0001-6850-1184 (M. C. Paoletti)

Table 1
ISO/IEC standard involved in AI

Field of application	ISO/IEC standards
Data quality for analytics and machine learning (ML)	ISO/IEC 5259-1:2024 ISO/IEC 5259-2:2024 ISO/IEC 5259-3:2024 ISO/IEC 5259-4:2024 ISO/IEC 5259-5:2024 ISO/IEC 25012:2008
AI Management system	ISO/IEC 42001:2023
AI concepts and terminology	ISO/IEC 22989:2022
Guidance on risk management	ISO/IEC 23894:2023
Controllability of automated artificial intelligence systems	ISO/IEC TS 8200:2024 ISO/IEC 25059:2023
Data life cycle framework	ISO/IEC 8183:2023
Requirements for bodies providing audit and certification of AI	ISO/IEC FDIS 42005
AI management systems and AI system impact assessment	ISO/IEC DIS 42006
Framework for Artificial Intelligence Systems Using Machine Learning	ISO/IEC 23053:2022
Ethical concerns	ISO/IEC TR 24368:2022 ISO/IEC 25059:2023

- ISO/IEC 22989 [6]: this standard provides a basic conceptual and terminological framework for artificial intelligence, improving cohesion between governance and interoperability components;
- ISO/IEC 23053 [7]: introduces a framework for AI systems using machine learning, outlining fundamental criteria for the development and responsible use of these technologies;
- ISO/IEC 23894 [8]: offers guidelines on managing the risks associated with the use of AI, which is critical to mitigating potential harms;
- ISO/IEC 25059 [9]: offers insight into future regulatory developments in the field of AI, outlining priorities for standardization;
- ISO/IEC FDIS 42005: deals with establishing a methodology for assessing the impact AI systems can have in various contexts. It covers aspects such as security, privacy, and ethical, social and legal implications. Impact assessment is critical to ensure that AI systems are developed so that risks are identified and mitigated before systems are implemented;
- ISO/IEC DIS 42006 [10]: establishes requirements for entities offering audit and certification services for artificial intelligence management systems, ensuring external quality control;
- ISO/IEC 8183 [11]: This standard addresses ethical and legal issues related to the use of AI, with a focus on transparency, accountability and privacy protection;
- ISO/IEC TS 8200 [12]: explores emerging techniques in AI, providing guidance for the safe and responsible adoption of new AI technologies.

In the area of data quality used in artificial intelligence and machine learning, we are helped by the ISO/IEC 5259 series. These standards are developed to ensure that the data used in analytics processes and machine learning models are of high quality so that the results produced are reliable, accurate and useful. The series consists of several parts, each of which addresses specific aspects of data quality:

- ISO/IEC 5259-1 [13]: This first part of the series provides a general introduction to data quality in AI and ML. It defines key concepts, terminology, and provides an overview of the main elements that influence data quality;
- ISO/IEC 5259-2 [14]: focuses on specific data quality measures, provides a model for data quality, and describes measures that can be used to assess data quality in analytics and machine learning contexts;

- ISO/IEC 5259-3 [15]: Provides requirements and guidelines for data quality management. It covers the practices and processes that organizations should implement to ensure that the data used in AI and ML applications are of high quality. This includes managing the data lifecycle, establishing data quality policies, and managing the risks associated with the data;
- ISO/IEC 5259-4 [16]: focuses on the specific processes required to maintain and improve data quality over time. This standard provides a structured framework of processes that organizations must follow to monitor and continuously improve data quality;
- ISO/IEC FDIS 5259-5 [17]: addresses data quality governance, i.e., the structures and control mechanisms that organizations should implement to ensure that data quality is effectively managed.

Data quality is critical because AI systems depend on data to learn, make decisions, and operate. If the data are inaccurate, incomplete or inconsistent, the decisions and results produced by AI can be equally flawed, with potential negative consequences.

ISO/IEC 5259-2 naturally intersects with other standards governing the management and implementation of artificial intelligence, particularly ISO/IEC 42001.

A key aspect of the connection between these two standards concerns data risk management. In fact, while on the one hand ISO/IEC 5259-2 emphasises the importance of continuous monitoring and improvement of data quality to prevent poor quality information (e.g. completeness) from compromising the validity of the AI model. On the other hand, the analysis of data risks is a topic closely related to the requirements of ISO/IEC 42001. The ISO/IEC 5259-2 also offers detailed guidelines for measuring and reporting data quality, aspects crucial to ensuring transparency and effective governance, central pillars of ISO/IEC 42001 as well.

Finally, we conclude our examination with a nod to CEN/CLC/TR 18115 "*Data Governance and quality for AI in the European context*", which is currently under publishing. It is intended to provide an overview of existing international and European regulations and standards.

3. Standard ISO and ethics in AI

The accelerating development of AI is opening up very relevant questions about how to address this technology in an ethical way, significantly affecting various areas of daily life, from privacy compliance to work management, from automation to health care, to the use of smart services, in politics and marketing. Wanting to delve into the ISO standards that deal with ethics, and considering the difficulties in defining unambiguous ethical directions for such complex and evolving technologies, it is useful to examine how the standards address these issues. ISO/IEC 25059 and ISO/IEC TR 24368 [18] provide a technical framework for mitigating ethical and social risks, focusing on transparency, accountability and quality, proposing models that address the challenges highlighted earlier.

ISO/IEC 25059 is part of the Systems and Software Quality Requirements and Evaluation (SQuaRE) series, developed by ISO/IEC JTC1/SC7. The SQuaRE series is focused on extending quality models to include specifications related to artificial intelligence systems. It introduces new characteristics and subcharacteristics for AI systems, such as transparency, controllability, and functional adaptability, all directly related to ethical challenges. Transparency, according to ISO/IEC 25059, implies that stakeholders should have access to detailed information about how an AI system works, including data about machine learning models.

Transparency: Both standards place transparency at the heart of the ethical discussion. While ISO/IEC 25059 describes it as a technical feature that ensures access to information about AI system processes, ISO/IEC TR 24368 expands it, linking it to the importance of ensuring that users understand and can trust automated decisions. This is a key element in preventing AI from becoming an incomprehensible "*black box*" with potential risks to human autonomy and dignity.

Controllability and accountability: The ability of human intervention on AI systems, guaranteed by the controllability described in ISO/IEC 25059, is essential to support the accountability principle discussed in ISO/IEC TR 24368. Both standards emphasize that without tools that allow humans to intervene in AI systems, it would be difficult to assign responsibility for errors or damage.

Non-discrimination and bias: The two standards also address the need to avoid discrimination and bias. ISO/IEC 25059 highlights the risks of bias in training data and the difficulty of achieving "*functional fairness*" in AI systems, while ISO/IEC TR 24368 delves into equity and social justice [19], proposing measures to ensure that AI does not reproduce or amplify existing biases.

In summary, the two standards provide a framework that delves into technical and social aspects of artificial intelligence. These documents provide practical guidance for developing ethical AI systems, promoting transparency, accountability, and the prevention of bias [5, 20], in line with the challenges and principles already outlined in the European AI Guidelines.

4. Conclusions

One of the most significant aspects that emerged concerns the need for a regulatory approach that is both integrated and flexible. Technical standards, such as ISO/IEC 42001 are essential to ensure the security and reliability of these technologies. However, it is important that there is integration in the areas of interoperability and governance (ISO/IEC 22989, ISO/IEC 23894). The adoption of quality standards on data quality has direct implications for the results of an AI system. Poor quality data poses a significant risk to the proper functioning of AI systems, leading to poor decisions that can compromise a company's reputation and legal compliance. As highlighted in [5, 21], data quality is a critical factor in avoiding biased results in AI systems. Through the adoption of combined metrics, such as those based on combinatorial calculus and frame theory, it is possible to preemptively assess the completeness and balance the data to mitigate such risks. The authors emphasize the need to develop an "*ethics-by-design*" approach, in which ethical considerations are incorporated from the ideation stages into the product life cycle. This approach is already present in ISO/IEC 42001 (Annex A) and can help prevent biases in AI systems, which could have discriminatory consequences and impacts on people. Integrating technical standards with ethical principles also helps build greater trust in AI systems. A key aspect for the adoption and success of new technologies, and the perception of individuals that they can be trusted, not only because there are norms that require it, but also because of an inherent assurance of the tool.

References

[1] ISO/IEC 42001 Information technology — Artificial intelligence — Management system, 2023. URL: https://www.iso.org/standard/44545.html.

[2] ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model, 2008. URL: https://www.iso.org/standard/35736.html.

[3] ISO 31000:2018 Risk management — Guidelines, 2018. URL: https://www.iso.org/standard/65694.html.

[4] A. Simonetta, A. Vetrò, M. C. Paoletti, M. Torchiano, Integrating square data quality model with iso 31000 risk management to measure and mitigate software bias, CEUR Workshop Proceedings 3114 (2021) pp. 17–22.

[5] A. Simonetta, M. C. Paoletti, A. Venticinque, The use of maximum completeness to estimate bias in AI-based recommendation systems, CEUR Workshop Proceedings 3360 (2022) pp. 76–84.

[6] ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, 2022. URL: https://www.iso.org/standard/74296.html.

[7] ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), 2022. URL: https://www.iso.org/standard/74438.html.

[8] ISO/IEC 23894 Information technology — Artificial intelligence — Guidance on risk management, 2023. URL: https://www.iso.org/standard/77304.html.

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

[10] ISO/IEC DIS 42006 Information technology — Artifcial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems, Under development. URL: https://www.iso.org/standard/44546.html.

[11] ISO/IEC 8183:2023 Information technology — Artificial intelligence — Data life cycle framework, 2023. URL: https://www.iso.org/standard/83002.html.

[12] ISO/IEC TS 8200:2024 Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems, 2024. URL: https://www.iso.org/standard/83012.html.

[13] ISO/IEC 5259-1:2024 Artificial intelligence — Data quality for analytics and machine learning (ML), 2024. URL: https://www.iso.org/standard/81088.html.

[14] ISO/IEC 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML), 2024. URL: https://www.iso.org/standard/81860.html.

[15] ISO/IEC 5259-3 Artificial intelligence — Data quality for analytics and machine learning (ML), 2024. URL: https://www.iso.org/standard/81092.html.

[16] ISO/IEC 5259-4 Artificial intelligence — Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework, 2024. URL: https://www.iso.org/standard/81093.html.

[17] ISO/IEC FDIS 5259-5 Artificial intelligence — Data quality for analytics and machine learning (ML) - Part 5: Data quality governance framework, Under development. URL: https://www.iso.org/standard/84150.html.

[18] ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concerns, 2022. URL: https://www.iso.org/standard/78507.html.

[19] J. Larson, S. Mattu, L. Kirchner, J. Angwin, Compas recidivism dataset, 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[20] A. Simonetta, T. Nakajima, M. C. Paoletti, A. Venticinque, Fairness metrics and maximum completeness for the prediction of discrimination, CEUR Workshop Proceedings 3356 (2022) pp. 13–20.

[21] A. Simonetta, A. Trenta, M. C. Paoletti, A. Vetrò, Metrics for identifying bias in datasets, SYSTEM 3118 (2021) pp. 10–17.