Responsible and Reliable AI @ CINI AI-IS

Piercosma Bisconti Lucidi[3], Lidia Marassi[1], Stefano Marrone[1,*], Domenico D. Bloisi[2], Daniele Nardi[3] and Carlo Sansone[1]

[1]*Department Of Electrical Engineering And Information Technologies, University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy*

[2]*Department of International Humanities and Social Sciences, International University of Rome, Via Cristoforo Colombo 200, Rome, 00147, Italy*

[3]*Department of Computer, Control and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, 00185, Italy*

Abstract

As Artificial Intelligence (AI) systems are increasingly deployed in critical sectors, ensuring their responsible and reliable operation has become a priority. This paper examines the importance of technical conformity verification as a mechanism for embedding principles of fairness, transparency, and safety within the AI development lifecycle. It distinguishes between commercial and research-grade systems, highlighting how their respective maturity levels influence the depth of conformity assessments. The study emphasizes the strategic placement of verification efforts after model evaluation and before deployment to mitigate risks related to bias, opacity, and lack of accountability. Drawing on international standards such as ISO/IEC TR 24027, the authors propose a structured approach to conformity verification that includes checklists and bias assessment protocols. The goal is to foster AI systems that are not only high-performing but also trustworthy, inclusive, and aligned with ethical norms. Future work will involve validating these practices through empirical case studies and fostering collaboration between academia, industry, and regulatory bodies.

Keywords

AI lifecycle, technical conformity, bias mitigation, ISO/IEC TR 24027, responsible AI, reliable AI, conformity assessment, AI standards

1. Introduction

Artificial Intelligence (AI) systems are being rapidly integrated into decision-making pipelines across diverse sectors, including healthcare, finance, transportation, and public administration. These technologies promise to enhance efficiency, reduce human error, and enable data-driven policy and operational decisions. However, despite their technical advancements, AI systems often carry risks related to bias, opacity, and limited accountability [1, 2]. These risks are frequently embedded in the development process itself, which typically follows four key stages: data collection and preparation, model development, evaluation, and deployment. Each phase of

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

piercosma.bisconti@consorzio-cini.it (P. B. Lucidi); lidia.marassi@unina.it (L. Marassi); stefano.marrone@unina.it (S. Marrone); domenico.bloisi@unint.eu (D. D. Bloisi); nardi@diag.uniroma1.it (D. Nardi); carlo.sansone@unina.it (C. Sansone)

0000-0001-8052-0142 (P. B. Lucidi); 0009-0006-8134-5466 (L. Marassi); 0000-0001-6852-0377 (S. Marrone); 0000-0003-0339-8651 (D. D. Bloisi); 0000-0001-6606-200X (D. Nardi); 0000-0002-8176-6950 (C. Sansone)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-07

this lifecycle introduces potential failure points. Biased training data can lead to discriminatory models [3], algorithmic design choices may amplify unintended patterns [2], and even comprehensive evaluations can fall short if they lack demographic disaggregation or ignore contextual variables [4]. Once deployed, these systems may be difficult to audit or correct, particularly in high-stakes or regulated environments.

In light of these challenges, there is growing consensus around the need for AI systems to conform to established *technical standards* that codify principles of fairness, transparency, and reliability [5]. These standards serve as a foundation for the responsible development of AI, providing normative guidance on risk assessment, bias mitigation, and system accountability. Among these, ISO/IEC TR 24027[1] offers specific recommendations for identifying and addressing bias across all phases of the AI lifecycle, from dataset construction to model deployment.

Yet the mere existence of such standards is not sufficient to ensure ethical or robust outcomes. To translate principles into practice, they must be operationalized through concrete *conformity assessment protocols*—structured procedures that systematically evaluate whether an AI system meets the requirements defined by the standards. These protocols should be tailored to different stages of technological maturity and integrated directly into the AI development workflow, allowing developers to anticipate and correct potential failures before systems reach production environments. As demonstrated by real-world incidents, such as the temporary suspension of ChatGPT in Italy due to regulatory concerns over data processing and transparency, neglecting these safeguards can have immediate legal and reputational consequences.

In this context, embedding conformity verification into the AI lifecycle is not merely a regulatory formality but a critical component of building AI systems that are not only performant but also fair, transparent, and trustworthy by design.

2. Technical Conformity Verification

Ensuring that AI systems meet established technical standards is essential for promoting safety, reliability, and public trust. This is particularly relevant for systems deployed in sensitive or high-impact environments. Technical conformity verification involves assessing whether an AI system adheres to normative requirements regarding its design, development, and behavior. This section outlines key considerations for integrating conformity assessment within the AI lifecycle, distinguishing between types of systems, and identifying the appropriate timing for such assessments.

2.1. Commercial vs. Research-Grade AI Systems

AI systems differ significantly in their readiness for deployment. For this reason, it is important to distinguish between *commercial products* and *research-grade systems*, as the requirements for conformity verification vary accordingly.

A widely adopted framework for measuring technological maturity is the Technology Readiness Level (TRL) scale. Originally developed by NASA, TRL levels range from early conceptual

[1] https://www.iso.org/standard/77608.html

stages (TRL 1–3) to fully operational systems (TRL 7–9). Research prototypes typically fall within TRL 1–6, while commercial products are generally situated at TRL 7 and above[2].

- **TRL 1–3:** Basic principles and proof-of-concept exploration, often limited to academic research.
- **TRL 4–6:** Prototype development and validation in relevant settings, representing the so-called "valley of death" in innovation.
- **TRL 7–9:** Final stages of testing, system qualification, and commercial deployment.

While full conformity verification is mandatory for commercial AI solutions, research-oriented systems may undergo a lighter form of evaluation, primarily focusing on methodological validity and reproducibility. Nonetheless, even research systems can benefit from partial conformity assessment, especially when addressing fairness, safety, or replicability in applied contexts.

2.2. Positioning Conformity Verification Within the AI Lifecycle

Conformity verification should be strategically integrated within the AI development process. The optimal point for intervention is after the model evaluation phase and before deployment. This allows for a comprehensive assessment of risks and limitations before the system is made available to end-users or integrated into operational environments.

Figure 1 illustrates the typical AI development pipeline and highlights where conformity assessment is most effectively applied.

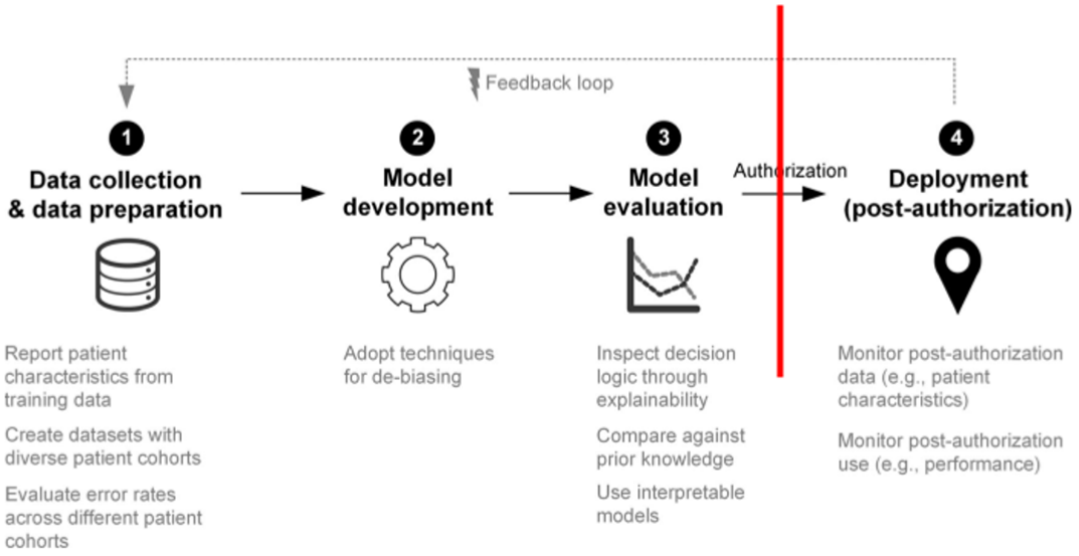

Figure 1: Positioning conformity verification in the AI development lifecycle (adapted from [6]).

[2]See NASA TRL definitions: https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html

Conducting verification before deployment is critical for minimizing legal, ethical, and reputational risks. Without it, issues such as inadequate bias control or insufficient data governance may only surface post-deployment, making remediation costly or infeasible. A notable example of premature deployment is the temporary suspension of ChatGPT in Italy due to concerns over user verification and data processing transparency[3]—demonstrating the high stakes involved in releasing unverified AI systems.

2.3. Why Verification Matters

A structured conformity process supports compliance with technical and ethical expectations, especially in applications involving sensitive user data or life-impacting decisions. In addition to addressing model performance and safety, conformity verification ensures:

- **User protection:** Verifying mechanisms for access control and risk mitigation.
- **Data privacy:** Assessing transparency in data usage and storage.
- **Bias detection:** Identifying and addressing sources of algorithmic unfairness, in line with standards such as ISO/IEC TR 24027[4].

In this context, standards play a foundational role. Yet their effectiveness depends on implementation through structured, repeatable processes. The next sections propose a framework to operationalize this verification through detailed checklists and bias assessments based on international technical guidelines.

3. Conclusion

As AI systems become increasingly embedded in real-world applications, ensuring their alignment with technical standards is not merely a regulatory formality—it is a necessary safeguard for ethical, reliable, and accountable deployment. The complexity and autonomy of these systems demand proactive mechanisms for risk identification, bias control, and performance validation, particularly in high-stakes domains where the consequences of technical failures may affect human rights, public safety, or institutional trust.

In this paper we outlined the rationale for integrating conformity verification into the AI development lifecycle, emphasizing the importance of distinguishing between research-grade prototypes and commercial-grade systems. Such differentiation enables the application of proportionate verification protocols tailored to the system's level of technological maturity and intended deployment context. Importantly, we have highlighted the optimal placement of conformity checks—between model evaluation and deployment—where they can most effectively prevent costly or irreparable failures in production settings.

Beyond assessing performance, conformity assessment frameworks grounded in international standards such as ISO/IEC TR 24027 provide essential tools for the systematic identification

[3]See official press release from the Italian Data Protection Authority: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847

[4]ISO/IEC TR 24027:2021, "Information technology — Artificial intelligence — Bias in AI systems and AI aided decision making," available at: https://www.iso.org/standard/77607.html

and mitigation of algorithmic bias, data governance weaknesses, and opaque engineering choices. By incorporating structured verification steps—such as bias audits, documentation reviews, and checklists—developers can embed fairness, transparency, and safety directly into the development process, rather than treating these as post-hoc concerns.

The need for such frameworks is particularly pressing in critical applications involving sensitive data or life-altering decisions. In these contexts, the absence of reliable conformity verification not only exposes users to harm but also undermines public confidence in AI systems more broadly.

Looking ahead, future work should aim to refine and validate these verification protocols through empirical case studies and cross-sectoral collaboration. Engagement between standards bodies, regulatory institutions, industry practitioners, and academic researchers will be crucial to ensure that verification practices remain up-to-date, context-sensitive, and practically implementable. Only through such integrated, interdisciplinary efforts can we build AI systems that are not only powerful, but also just, inclusive, and trustworthy by design.

Declaration on Generative AI

During the preparation of this work, the authors used GPT and DeepL to perform grammar and spelling check. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (2021).

[2] H. Suresh, J. V. Guttag, A framework for understanding unintended consequences of machine learning, Communications of the ACM 64 (2021) 62–71.

[3] S. Barocas, A. D. Selbst, Big data's disparate impact, California Law Review 104 (2016) 671–732.

[4] B. Mittelstadt, Principles alone cannot guarantee ethical ai, Nature Machine Intelligence 1 (2019) 501–507.

[5] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399.

[6] K. N. Vokinger, S. Feuerriegel, A. S. Kesselheim, Mitigating bias in machine learning for medicine, Communications medicine 1 (2021) 25.