Journal of Electronic Research and Application, 2026, Volume 10, Issue 9
https://bbwpublisher.com/journal/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Rationale and Governance Framework for the Prudent Adoption of Artificial Intelligence in Traditional Manufacturing Enterprises

Xin He, Xuejiao He, Yuhan Wu*

Changhe Aircraft Industries (Group) Co., Ltd., Jingdezhen 333000, Jiangxi, China

***Corresponding author:** Yuhan Wu, wuyh074@avic.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Artificial intelligence (AI) is permeating traditional manufacturing through applications such as production scheduling, quality inspection, equipment monitoring, supply chain collaboration, and knowledge retrieval. This paper analyzes the practical drivers behind manufacturers' adoption of AI, proposes a governance framework for prudent AI adoption built on three mechanisms: a risk-grading and determination process, a mandatory human-intervention protocol, and a soft-fault protection layer design. These three mechanisms correspond respectively to task selection (which tasks are suitable for AI), operational oversight (who retains oversight during operation), and failure response (how to respond when AI fails), forming a closed loop that spans pre-deployment assessment, in-operation supervision, and post-failure response. The role of AI in manufacturing should follow two principles: boundary limitation (AI undertakes critical tasks only when boundaries are clear, performance validated, and independent safety layers exist) and role restriction (in high-risk scenarios, AI serves as an assistive tool rather than an exclusive controller).

Keywords: Artificial intelligence; Traditional manufacturing; Risk-grading; Human intervention; Soft-fault protection; Governance framework

Online publication: October 30, 2026

1.Introduction

Manufacturing systems demand high levels of information integration and process control. Although traditional manufacturing enterprises possess substantial process expertise and stable customer relationships, their operational foundations are often characterized by fragmented data, heterogeneous equipment, incomplete historical records, and experience-dependent decision-making. Fluctuating demand, unstable supply chains, tightening safety requirements, and generational workforce transitions have further widened this gap, making it increasingly urgent to stay abreast of production status and operational risks [1].

AI can handle repetitive, structured tasks, but it cannot replace professional judgment. Two approaches

are particularly problematic: treating AI merely as an extension of conventional automation, or rushing to deploy it without first defining business requirements and acceptance criteria. Avoiding these extremes requires a governance approach that embeds AI adoption within a disciplined framework—one that spans the entire system life cycle [2].

This paper proposes three mechanisms corresponding to the three stages of "selection–control–protection" (**Figure 1**). The first is a risk-grading and determination process, which classifies AI tasks into low-, medium-, and high-risk levels and assigns differentiated control requirements. The second is a mandatory human-intervention protocol, specifying mandatory confirmation conditions, operator veto authority, and automatic degradation logic. The third is a soft-fault protection layer design, which extends trigger signals from physical limit violations to model state anomalies and responds through dynamic safety envelopes and systematic degradation. Together, these mechanisms constitute a complete governance framework spanning task assessment, operational control, and failure response, translating principle-based guidance into actionable manufacturing engineering requirements.

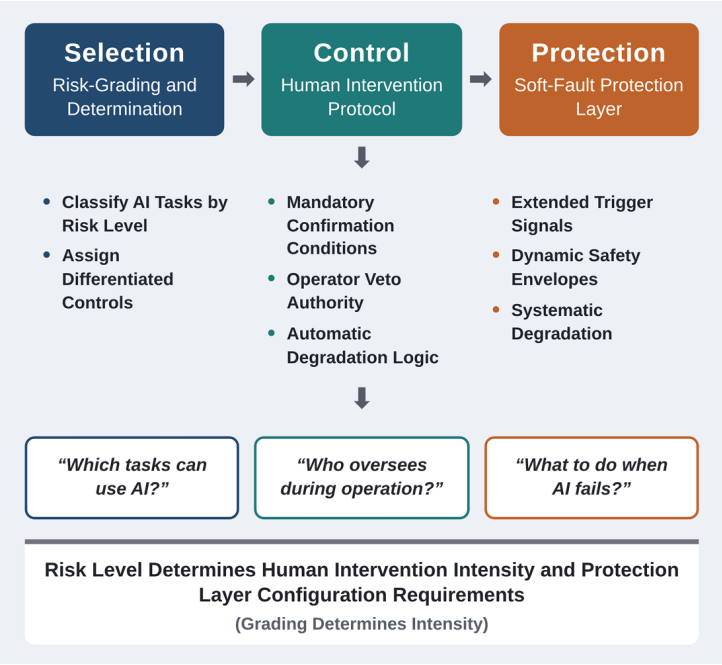


Figure 1. Progressive relationship and closed-loop governance of the three mechanisms.

2.Risk-Grading and Determination Process

This section presents the complete construction of the risk-grading and determination process. It first provides a general risk-grading methodology (**Table 1** and **Figure 2**), then demonstrates its application through seven typical manufacturing AI tasks, and finally summarizes the grading results (**Table 2** and **Figure 3**).

Table 1. Risk-grading questionnaire

Dimension	Question	Low-risk reference	Medium-risk reference	High-risk reference
Consequence severity	What is the most serious consequence if the AI output is erroneous?	Misinformation or efficiency loss	Quality defects or equipment damage	Personal injury
Reversibility	Can an error be detected and corrected before it causes substantial harm?	Error is reversible	Error can be detected at the next stage with controllable cost	Error is irreversible and causes immediate loss
Feasibility of human verification	Can the AI output be independently verified by a human?	Yes	Theoretically possible, but practically difficult	No
Match between AI capability and task requirements	Does the task demand deterministic responses, explainability, or the ability to handle edge cases? Can AI meet these demands?	Complete match	Partial match	Fundamental mismatch

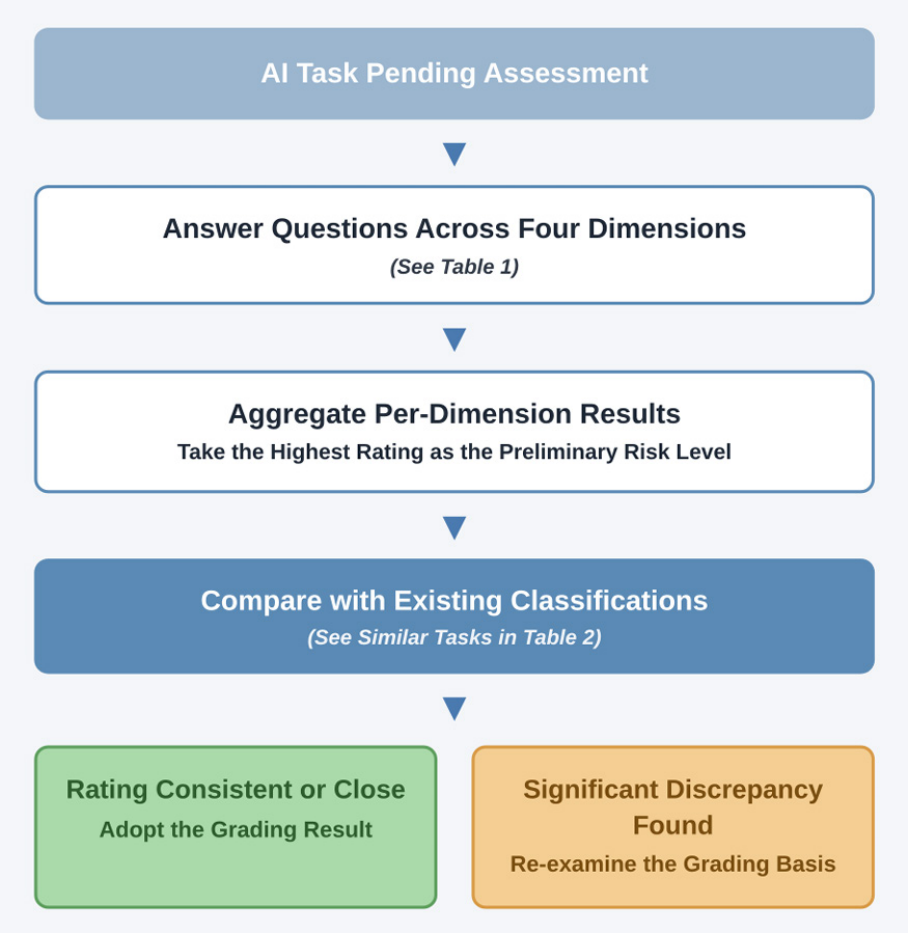


Figure 2. Risk-grading decision workflow.

Table 2. Risk classification and control requirements for seven typical manufacturing AI tasks

Task	Description	Risk level	Validation requirement	Human intervention intensity	Execution authority

Knowledge retrieval	Retrieval of approved technical documents and maintenance records	Low	Compliance check of document sources	User self-check	Results may be returned directly but are not to be used as operational instructions
Production scheduling assistance	Generation of recommended schedules based on historical orders and inventory	Low	Offline back-testing	Optional review	Recommendations only; changes require planner confirmation
Equipment anomaly indication	Alerts on equipment state deviations based on temperature, vibration, and other signals	Medium	Representative operating condition replay validation	Maintenance engineer confirmation	Alarms only; no shutdown actions triggered
Demand forecasting	Assisted demand forecasting based on historical orders and inventory	Medium	Rolling forecast error assessment	Planner review	Forecasts only; order adjustments require approval
Visual inspection (assistive)	AI pre-screening of product surface defects with annotation of suspect areas	Medium	Online small-batch validation and false detection rate monitoring	Human review of suspect areas	Screening and alerts only; no direct pass/fail decisions
Quality control release	Automated determination of product release based on inspection results	High	Real-world operating condition validation and independent review	Mandatory human confirmation	No automated release execution
Safety protection functions	Safety interlock and emergency stop decisions	High	Functional safety verification (SIL assessment)	Independent protection layer priority	Not to be performed by AI independently



Figure 3. Overview of risk classification for typical manufacturing AI tasks.

2.1.Risk-grading questionnaire and decision workflow

To facilitate enterprises' self-assessment of AI task risk levels, this paper develops a structured risk-grading questionnaire (**Table 1**) and an associated decision workflow (**Figure 2**). Enterprises follow the workflow to answer questions across four dimensions, aggregate the results (low/medium/high), and take the highest rating as the preliminary risk level for the task.

2.2.Risk-grading of seven typical tasks

Knowledge retrieval indexes only approved documents without generating new instructions, and is classified as low risk. Production scheduling assistance generates only recommendations, with the final schedule taking effect only after planner confirmation; it is also classified as low risk.

Equipment anomaly indication identifies statistical deviations rather than physical causes, with alarms requiring diagnostic confirmation by maintenance engineers before action; it is classified as medium risk [3]. Demand forecasting produces probabilistic predictions, with planner review as a mandatory step before order adjustments; it is classified as medium risk. Visual inspection can perform initial screening, but "reject" decisions require manual confirmation by quality inspectors; it is classified as medium risk [4].

Quality control release decisions are irreversible, and AI should not independently perform release judgments; they are classified as high risk. Safety protection functions demand deterministic responses, which are structurally incompatible with AI's probabilistic outputs; they are also classified as high risk.

2.3.Risk-grading overview

Based on the analysis in Section 2.2, the risk-grading results for each task and their distribution across the risk spectrum are illustrated in **Figure 3**.

3.Design basis for the human-intervention protocol and soft-fault protection layer

The risk-grading process defines AI's operational boundaries, but two further questions arise: how to ensure timely human intervention when AI makes errors, and how the protection layer can perceive and respond when AI fails without conventional alarms being generated. This section analyzes five categories of common risks that inform the mandatory confirmation conditions in the human-intervention protocol and underpin the trigger-signal expansion and dynamic safety envelope design in the soft-fault protection layer.

3.1. Data quality risk

Industrial data rarely possess ideal learning conditions. Missing records, incorrect labels, imbalanced sampling, class imbalance, and changes in recording practices can all cause learned relationships to break down in operation. The problems are systemic and insidious: sensor drift and calibration bias shift inputs away from training baselines; industrial labeling often relies on inspection records or logs as "ground truth" that may themselves contain errors; normal operating data dominate while critical transient data (startups, shutdowns, changeovers, faults) are scarce; safety incident rarity exacerbates positive-sample shortage; and equipment replacement or line modifications can radically alter data distributions (concept drift [5]). Enterprises should therefore systematically evaluate data relevance, representativeness, provenance, and completeness rather than using volume as the sole metric, and establish data health checks to support the confirmation conditions in the human-intervention protocol.

3.2. Inherent risks of AI models

Even with reliable data, AI models introduce probabilistic outputs, performance drift, and lack of explainability. Test-stage accuracy does not guarantee long-term production reliability; changes in raw materials, specifications, equipment condition, lighting, or work practices can degrade performance. Critically, AI "soft faults" differ fundamentally from conventional failures (**Table 3**): traditional fail-safe designs assume visible, sensor-capturable failures (motor stoppage, pressure overshoot, excessive temperature), whereas AI continues running and producing outputs whose quality degrades imperceptibly, generating no signals that conventional sensors or limit checks can capture. Overall accuracy may mask consequential errors; safety- and quality-critical applications should separately evaluate miss rates, false-positive rates, rare-case performance, anomalous-input response, and boundary-condition stability, linking each metric to its corresponding operational cost, with risk management continuing after deployment [2,6].

Table 3. Comparative risk profiles: traditional automation systems vs. AI systems

Dimension	Conventional deterministic automation systems	AI systems
Operating characteristics	Operates on deterministic logic or analytical rules; behavior is predictable; boundaries are fixed.	Probabilistic outputs, model drift, and lack of explainability; highly complex decision boundaries.
Failure mode	Hard faults—known hardware failures or parameter limit violations, following fixed failure paths.	Soft faults—error patterns evolve dynamically with input data distribution changes; no fixed failure path.
Failure manifestation	Shutdown—clear physical signals such as motor stoppage, pressure overshoot, or excessive temperature.	Silent degradation—the model continues to run and produce outputs, but output quality degrades imperceptibly.
Detectability	Failures are visible and can be captured by sensors or limit checks.	Does not manifest as physical quantity limit violations; cannot be captured by conventional sensors or limit checks.
Response strategy	Pre-engineered strategies exist: redundancy, interlocks, and watchdog timers.	New response mechanisms must be designed: continuous monitoring of fine-grained metrics and ongoing post-deployment risk management.

Core difference: AI's failure mode is not "shutdown" but "silent degradation"—traditional protection layers only need to respond to clear physical signals, whereas AI "soft faults" require new detection and response mechanisms.
Fine-grained metrics for continuous monitoring: miss rate, false-positive rate, performance on rare cases, response to anomalous inputs, and stability under boundary conditions, all linked to their corresponding operational costs.

3.3. Interface risk

When an AI model is embedded in an existing manufacturing system, new risks emerge at the interaction boundary between data inputs and command outputs. Integration testing can hardly exhaust all interaction paths; some error combinations only manifest under specific production conditions. To address the six categories of interface risks (**Figure 4**), the system should adopt four protective measures: input validation (format and range checks, with nonconforming inputs triggering alarms), output limiting (configurable upper/lower thresholds with automatic blocking), exception handling (predefined fallback modes for AI unavailability), and human takeover procedures (maintained intervention channels under abnormal conditions). Opaque or insufficiently validated models should not be entrusted with safety protection functions independently [7].

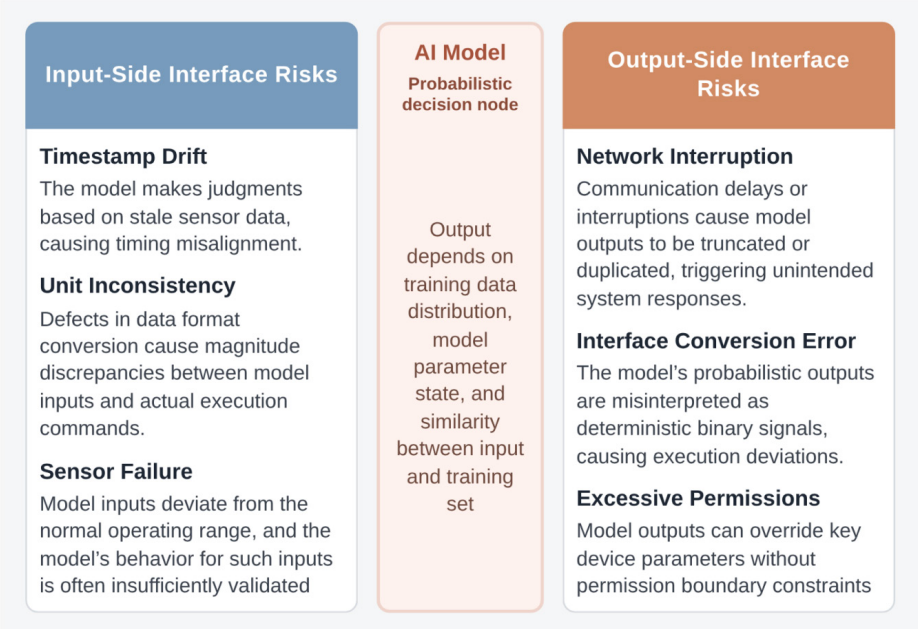


Figure 4. Interface risks in AI-integrated automation systems.

3.4. Cybersecurity risk

AI significantly expands the network attack surface: production systems originally operated in isolated industrial networks, but AI applications require connections to external data sources, cloud services, and remote maintenance, introducing new attack vectors. Attackers may steal sensitive information (process parameters, design drawings, customer orders) or interfere with AI judgments through data poisoning or input tampering. Industrial data are simultaneously training data, inference inputs, and trade secrets; leakage not only affects competitiveness but also provides attackers with prior knowledge to craft adversarial examples. While enterprises can implement authentication, encryption, network segmentation, and log auditing internally, effectiveness depends on supplier cooperation in data retention, secondary training, cross-border transfers, and deletion completeness. Cybersecurity controls should therefore be considered at the architecture design and procurement stages rather than retrofitted after deployment [8].

3.5. Internal risk

Cybersecurity addresses external threats; internal risk addresses an enterprise's own governance deficiencies. Together they constitute two sides of accountability in AI adoption. The "black-box" nature of AI complicates accountability: when AI intervenes in decision-making, outputs are probabilistic results jointly determined by training data, model parameters, and algorithmic logic, making post-incident responsibility difficult to assign without model versions, operation records, and complete logs. Responsibilities should be allocated in advance among data providers (annotation quality), model developers (validation completeness), system operators (deployment and monitoring reliability), and business reviewers/final decision-makers (reasonableness of adopting recommendations). Traceability is the technical foundation: important outputs should be traceable to data sources, model versions, generation times, operators, and triggered actions; high-risk applications should establish formal approval, review, escalation, suspension, and auditing procedures to enable decision-chain reconstruction and learning from failures [6].

4. Implementation of the human-intervention protocol and soft-fault protection layer

This section addresses how the system ensures that AI errors are detected promptly, intercepted appropriately, and safely contained, elaborating on the human-intervention protocol and soft-fault protection layer, supplemented by continuous monitoring and traceability, as well as a preset exit mechanism.

4.1. Establishing the value premise of AI applications

The effective operation of the three mechanisms depends first on answers to two questions: "Does this task truly require AI?" and "Is this task suitable for AI?" If poor performance stems from disorganized processes, aging equipment, missing standards, or unclear responsibilities, basic management improvements often deliver greater value. AI introduction is justified only when processes are standardized, data quality is controllable, repetitive tasks have clear evaluation criteria, and manual processing faces efficiency or consistency bottlenecks. Each project should establish baselines, target values, costs, evaluation periods, and termination conditions in advance, since "becoming intelligent" is not itself a testable objective. This approach is consistent with the principle that smart manufacturing capabilities should develop progressively according to actual conditions and business needs [9].

4.2. Human-intervention protocol

The protocol establishes mandatory human-in-the-loop checkpoints during AI operation, consisting of three components: mandatory confirmation, veto authority, and automatic degradation, forming a closed loop of "detect–halt–respond."

Four conditions trigger mandatory confirmation: (1) output confidence falls below a preset threshold; (2) input data fall outside the training distribution (new products, materials, processes); (3) the model repeatedly triggers high-risk alarms or makes contradictory judgments; (4) output approaches or exceeds the safe operating boundary. When any condition triggers, the AI output must be human-confirmed before execution; conditions (1)–(2) address data quality risks, (3) addresses model drift, and (4) addresses interface risk.

Veto authority freezes execution authority when confirmation is triggered, giving operators full veto power whose decisions are recorded as binding instructions. Automatic degradation responds to consecutive verifiable errors by lowering the automation level (e.g., from "AI recommendation + human confirmation" to "manual operation only"), with manual reset and root-cause analysis required before restoration, and termination if fundamental defects are found. Because AI soft faults do not trigger explicit alarms, degradation logic automatically transforms the "model unreliable" state into a safety response without waiting for human judgment. Effective exercise of veto authority requires substantive human supervision: reviewers need process knowledge and training, and the system interface must provide sufficient evidence for independent judgment.

4.3. Soft-fault protection layer

Traditional protection layers target "hard faults" (motor stoppage, pressure overshoot, signal loss) that produce clear physical signals. AI's "soft fault" mode—continued operation with imperceptibly degrading output quality—generates no recognizable alarm signals, requiring redesigned perception and response through three adjustments.

First, trigger signal extension: beyond physical limit violations, the layer responds to model state

anomalies including sustained low confidence, input features deviating beyond allowable distances from the training distribution, and abnormal fluctuations in output temporal stability—intervening at the "judgment may be erroneous" stage rather than waiting for errors to occur. Second, dynamic safety envelopes: as input features approach training distribution edges and AI credibility decreases, the envelope automatically narrows, reducing AI's permissible output range in real time (e.g., restricting output to half the normal range when confidence falls below 80% of the training-set average). Third, elevation of degradation logic to system baseline: analogous to hardwired emergency stops, degradation is not an optional add-on but a fundamental design component, with each level's behaviors (retained/disabled functions, escalated alarms) and fallback paths predefined at the design stage.

The soft-fault protection layer should satisfy four requirements: (1) a monitoring channel independent of the AI model for detecting distribution shifts and confidence anomalies; (2) a dynamic safety envelope narrowing operational space by input features and model state; (3) predefined multi-level degradation states and corresponding behaviors; (4) automatic transfer of execution authority to backup mechanisms upon degradation until root-cause analysis is complete.

4.4. Continuous monitoring and traceability

Deployment is not the endpoint. Continuous monitoring covers performance metrics (false-positive rates, miss rates, input distribution changes, triggering alarms or re-validation when abnormal) and behavioral logs (confidence levels, trigger conditions, intervention records, execution results for each output). Important outputs should be traceable to data sources, model versions, generation times, operators, and triggered actions, enabling decision-chain reconstruction after incidents. Monitoring should be institutionalized through regular management reviews covering incidents, near-misses, false positives, missed detections, override records, degradation triggers, cybersecurity events, model changes, realized benefits, and maintenance costs, with the data feedback supporting model state awareness in the soft-fault protection layer.

4.5. Preset exit mechanism

The exit mechanism is the final safeguard, ensuring the enterprise can safely withdraw without weakening existing controls. It should be considered at project outset, with smooth fallback to existing processes when AI performance consistently fails and cannot be improved, risks escalate (new vulnerabilities or failure modes), business conditions change significantly, or superior alternatives emerge. Feasibility depends on binding supplier governance clauses specifying performance metrics, data ownership, model update strategies, audit cooperation, interface access, and exit conditions; enterprises must retain access to core data, configurations, logs, and documentation to avoid losing control of critical assets. If data integrity and accessibility cannot be maintained after supplier exit, the mechanism itself is unenforceable.

5. Implementation of the governance framework

A critical gap exists between theoretical framework and actual deployment: how mechanisms are activated sequentially and implemented in stages. This section embeds the three mechanisms into a six-step implementation process, enabling incremental rather than one-time deployment.

Step 1: Joint assessment. Engage stakeholders from business, equipment, data, personnel, cybersecurity, and safety to assess current conditions and pain points, identifying root causes and scope; the core output is

clarifying the most pressing issues and which areas possess AI adoption conditions. Step 2: Pilot selection. Choose scenarios with complete data, clear boundaries, manageable risk, and quantifiable metrics (e.g., assistive inspection for a product family or condition monitoring for selected equipment), corresponding to the first mechanism's selection function using **Table 1**. Step 3: Baseline establishment. Build a performance baseline using conventional methods before AI introduction, evaluating model metrics separately from business metrics and avoiding acceptance based solely on supplier demonstrations; baseline data are the sole reference for judging AI improvement. Step 4: Limited trial operation. Conduct small-scale trials while retaining manual processes, without altering critical control parameters initially; medium- and high-risk applications require human confirmation, corresponding to the second mechanism's operational configuration. Step 5: Revision and refinement. Based on trial data (false-positive rates, miss rates, intervention frequency, degradation triggers), revise operational procedures, permission settings, training content, and abnormal response plans. Step 6: Expansion decision. A joint review by business, technical, equipment, quality, and safety stakeholders decides whether expansion conditions are met, corresponding to the third mechanism's resilience test verifying independent protection layer takeover after AI failure; expansion proceeds only after confirmation.

Pilot acceptance should address: whether false alarms and missed detections are within acceptable limits; whether the system detects and reports data interruptions or anomalous inputs; whether users can understand outputs and conduct independent reviews; whether operations can be promptly switched to manual after failure; whether measurable benefits justify long-term maintenance and governance costs; and whether the supplier commits to interfaces, logs, documentation, and exit support. If any question remains unsatisfactorily answered, high-risk deployment should be deferred. A successful pilot does not automatically authorize full-scale rollout; conditions should be re-evaluated before expansion, and changes in products, equipment, materials, or environment should trigger a new validation round.

6. Conclusion

This paper provides a governance framework for the prudent adoption of AI in traditional manufacturing enterprises, built on three mechanisms: a risk-grading and determination process (defining AI's scope at the source), a mandatory human-intervention protocol (establishing human oversight checkpoints during operation), and a soft-fault protection layer design (providing an independent safety barrier at failure). Together they correspond to "selection," "control," and "protection," forming a complete governance loop covering pre-event, in-operation, and post-event phases. The framework is grounded in an analysis of five categories of common risks—data quality, inherent AI model, interface, cybersecurity, and internal risks—and operationalized through a six-step implementation pathway from joint assessment to expansion decision.

The framework is principles-based rather than a technical implementation solution: specific parameters (confidence thresholds, envelope tightening degrees, degradation-triggering error counts) require calibration in real operating contexts. Furthermore, the risk categories are based on seven typical manufacturing AI tasks; emerging scenarios beyond this scope (e.g., generative AI in manufacturing) may require further investigation of their risk characteristics.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Tilbury D, 2019, Cyber-Physical Manufacturing Systems. Annual Review of Control, Robotics, and Autonomous Systems, 2: 427–443.

[2] Tabassi E, 2023, Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology.

[3] Abidi M, Mohammed M, Alkhalefah H, 2022, Predictive Maintenance Planning for Industry 4.0 Using Machine Learning for Sustainable Manufacturing. Sustainability, 14(6): 3387.

[4] Hoffmann R, Reich C, 2023, A Systematic Literature Review on Artificial Intelligence and Explainable Artificial Intelligence for Visual Quality Assurance in Manufacturing. Electronics, 12(22): 4572.

[5] Lu J, Liu A, Dong F, et al., 2019, Learning Under Concept Drift: A Review. IEEE Transactions on Knowledge and Data Engineering, 31(12): 2346–2363.

[6] ISO/IEC 23894:2023, 2023, Information Technology—Artificial Intelligence—Guidance on Risk Management. Geneva: International Organization for Standardization.

[7] ISO/IEC TR 5469:2024, 2024, Artificial Intelligence—Functional Safety and AI Systems. Geneva: International Organization for Standardization.

[8] NIST, 2025, Cybersecurity Framework (CSF) 2.0 Manufacturing Profile. NIST IR 8183 Rev. 2 (Initial Public Draft). Gaithersburg, MD: National Institute of Standards and Technology.

[9] Monostori L, Kádár B, Bauernhansl T, et al., 2016, Cyber-Physical Systems in Manufacturing. CIRP Annals, 65(2): 621–641.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.