GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Decision Rights and Escalation Design in Industry 4.0: A Framework for Explainable Autonomous Operations

Suzhen Huang [1,*]

[1] Independent Researcher, China
* Correspondence: Suzhen Huang, Independent Researcher, China

Abstract: Industry 4.0 has made autonomous and semi-autonomous operations technically feasible, but many industrial studies still describe autonomy as if it were only a question of algorithms and data. In practice, industrial autonomy is a governance problem as much as a technical one. Managers need to specify who may decide, under what conditions machine recommendations become executable actions, how exceptions are escalated, and how learning is retained without weakening accountability. This paper develops a conceptual framework for decision rights and escalation design in Industry 4.0. The central argument is that industrial autonomy should be understood as a managed distribution of observation, interpretation, recommendation, authorization, execution, and learning rights rather than as a binary choice between manual and automatic operation. The paper organizes this argument through three linked ideas. First, autonomy should be calibrated by consequence severity, reversibility, and evidence quality, not by model confidence alone. Second, explainability and auditability are operational features that make autonomous actions governable rather than optional extras. Third, escalation design is the mechanism that preserves human control while still allowing fast and scalable machine-supported action. The framework is illustrated with examples from quality control, production scheduling, resource dispatching, service operations, and multi-state industrial systems. The paper contributes a governance-oriented interpretation of Industry 4.0 autonomy that is distinct from purely technical discussions and offers a practical research agenda for explainable and accountable industrial operations.

Keywords: industry 4.0; autonomous operations; decision rights; escalation design; explainability; human-in-the-loop

Received: 24 March 2026
Revised: 14 May 2026
Accepted: 28 May 2026
Published: 01 June 2026

1. Introduction

Industry 4.0 is frequently introduced through a well-known set of technologies, including cyber-physical systems, smart sensors, machine learning, industrial platforms, edge computing, and increasingly autonomous control. However, organizations often face challenges that go beyond merely acquiring data. The real difficulty lies in determining which actions machines are permitted to take independently, which actions require human oversight, and how automated recommendations can be trusted in situations involving significant costs or uncertainties [1]. These considerations are critical for ensuring that automation aligns with organizational goals and operational realities.

This distinction is significant because many industrial decisions fall within a spectrum between complete manual control and full automation. For instance, a scheduling engine may generate a daily plan, but supervisors might still need to approve overtime or decide to outsource certain orders [2]. Similarly, a vision model may identify potential defects, but operators are tasked with deciding whether to halt production. In another example, a field-service platform may automatically prioritize dispatches, yet managers may intervene during emergencies or resource shortages. Autonomy, therefore, is not a binary state but rather a nuanced distribution of decision-making rights across various levels of an organization. This layered approach allows for flexibility and adaptability in managing complex industrial processes.

The existing body of literature provides numerous valuable insights, but these insights are often fragmented across different domains [3]. Research on human-automation interaction explores concepts such as levels of automation, trust, and situational awareness. Studies on responsible AI emphasize the importance of explanation, audit mechanisms, and accountability frameworks. Meanwhile, industrial decision-making research highlights the challenges posed by partially observed states, interconnected resources, and actions with delayed consequences. Despite these contributions, there remains a need for a comprehensive framework that integrates these diverse strands into a unified decision-rights architecture tailored for Industry 4.0 autonomy. Such an architecture would provide organizations with a clear roadmap for balancing automation and human oversight effectively.

This paper aims to address this gap by treating Industry 4.0 autonomy as a governance challenge supported by a robust technical foundation. The central question is not merely whether a model possesses sufficient computational power to act autonomously, but rather how an organization can systematically structure the transition from automated recommendations to actionable decisions. The focus is placed on key elements such as decision rights allocation, calibration of autonomy levels, mechanisms for explanation, protocols for handling exceptions, and the establishment of learning boundaries. While the discussion is conceptual in nature, it remains closely tied to practical industrial tasks to ensure that the proposed framework is both applicable and operationally relevant. Figure 1 illustrates the decision-rights ladder for Industry 4.0 autonomy, providing a visual representation of the hierarchical distribution of decision-making authority [4, 5].

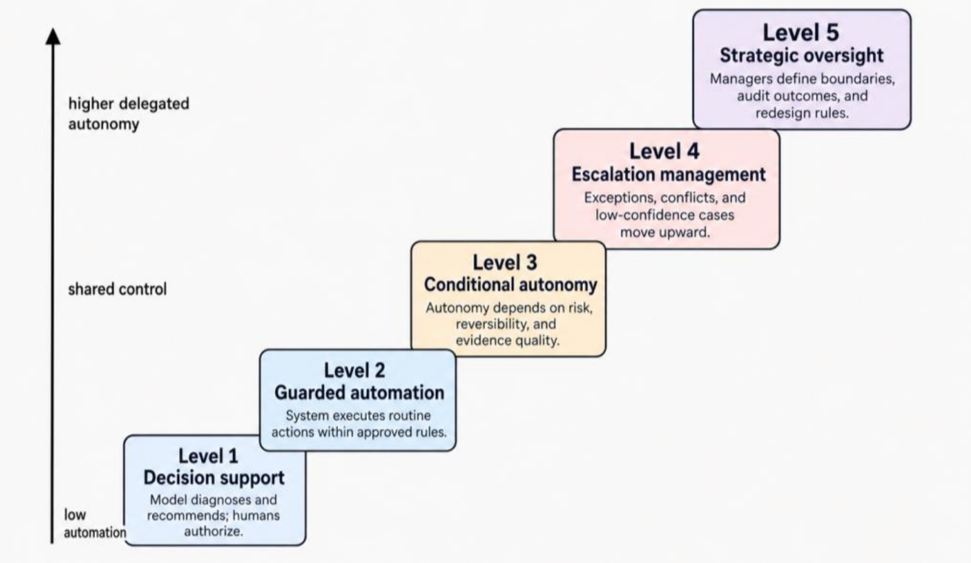

Figure 1. Decision-rights Ladder for Industry 4.0 Autonomy.

2. From Algorithmic Capability to Decision Rights

A useful starting point is to distinguish between algorithmic capability and institutional authority. While a model may excel at detecting patterns, forecasting outcomes, or ranking action alternatives, these technical capabilities alone do not determine who holds the authority to make decisions. Industrial autonomy becomes significant only when an organization systematically allocates a sequence of decision rights. These rights include the right to observe data, interpret findings, recommend actions, authorize decisions, execute tasks, and update the rule base after outcomes are evaluated [6]. This structured allocation ensures that the technical capabilities of algorithms are effectively integrated into the broader organizational framework, enabling meaningful and accountable decision-making processes.

Adopting a rights-based perspective fundamentally alters the design logic of automated systems. It helps avoid the common misconception that automation is merely a direct substitution of human labor with machine output. In practice, various actors often retain distinct rights simultaneously [7]. For instance, sensors and analytical models may dominate the processes of observation and interpretation, while supervisors might retain the authority to authorize decisions. Similarly, operators may exercise discretion in executing tasks within specific local contexts, and governance teams may retain the responsibility to revise thresholds and escalation rules following audits. This layered approach ensures that automation is not only technically efficient but also adaptable to the complexities of real-world organizational dynamics.

This decomposition of decision rights also provides a clearer framework for identifying and addressing failures. Errors can arise at multiple stages, such as during sensing, state estimation, recommendation ranking, or boundary management. For example, a policy might be mathematically sound but operationally infeasible if it violates an implicit constraint within the system [8]. Similarly, a recommendation might be statistically accurate but impractical if it cannot be communicated effectively to frontline personnel for timely adoption. By linking technical accuracy with organizational usability, decision-rights design ensures that automated systems are not only precise but also practical and actionable within their intended contexts.

Concrete industrial examples illustrate the importance of decision-right allocation. In service systems with multiple task categories, the critical question is not merely how to optimize a control rule but also who should have the authority to override it during periods of high demand or resource scarcity. In multi-component systems, decision-making may depend on hidden states and protective structures, necessitating organizational clarity on whether local operators can act based on summarized indicators or if approval must be escalated to higher levels. In multi-agent environments, additional complexities arise, such as how local policies should coordinate and what mechanisms should be in place to resolve disagreements among agents. These considerations highlight the need for a robust framework that balances technical efficiency with organizational adaptability (As shown in Table 1).

Table 1. Decision-right Decomposition for Industry 4.0 Autonomous Operations.

Decision right	Core question	Typical industrial owner	Governance issue
Observation	What is seen and recorded?	Sensors / operators	Data quality and coverage
Interpretation	How is state inferred?	Analytics / digital model	Model validity and assumptions
Recommendation	Which action is preferred?	Decision engine	Objective alignment and constraints
Authorization	Who approves action?	Supervisor / manager	Risk ownership and accountability
Execution	How is action implemented?	Operations / automation	Procedural compliance
Learning	What changes after outcomes?	Governance owner	Drift control and rule revision

3. Autonomy Calibration by Impact, Reversibility and Evidence Quality

Once decision rights are explicitly defined, the subsequent challenge lies in determining the appropriate level of machine authority for various decisions. Not all decisions warrant the same degree of autonomy, as the level of machine involvement should be calibrated based on several critical factors [9, 10]. These include the potential impact of an error, the reversibility of the action, and the quality of the evidence available to support the decision. While confidence scores generated by algorithms are important, they represent only one aspect of this broader calibration framework. A comprehensive approach is necessary to ensure that machine autonomy is applied judiciously and effectively.

Decisions involving low-impact and easily reversible actions are ideal candidates for routine autonomy. For instance, tasks such as minor sequencing adjustments, automated replenishment suggestions, or noncritical parameter tuning can often be delegated to machines with minimal oversight. On the other hand, decisions that carry significant consequences and are difficult to reverse typically require explicit human authorization. Examples of such high-stakes decisions include major schedule disruptions, switching suppliers during critical shortages, overriding safety protocols, or shutting down essential production capacities. Between these extremes lies a spectrum of guarded autonomy, where machines are permitted to act only within a predefined set of rules and constraints. This middle ground ensures a balance between efficiency and control, allowing for some degree of automation while maintaining safeguards against potential risks.

The quality of evidence further influences the calibration of autonomy. In scenarios where the impact of a decision is low but the evidence is unreliable or incomplete, a quick human review may still be necessary to ensure sound judgment. Conversely, in cases where the consequences are significant but the evidence is robust and procedures are highly standardized, a system of guarded automation may suffice, reducing the need for full manual intervention. This highlights the complexity of designing automation frameworks, as a single rule cannot adequately address the diverse range of decision-making scenarios. Instead, a structured matrix is required, one that systematically evaluates the interplay between consequences, reversibility, and evidence quality for each category of decisions. Such a matrix provides a nuanced approach to determining the appropriate level of machine autonomy [11].

This rationale underscores the continued relevance of advanced industrial optimization models in discussions about governance and automation. These sophisticated models offer valuable insights into the underlying structure of decision-making environments. They help identify critical factors such as resource interdependencies, hidden system states, and actions that may lead to delayed consequences. Understanding these structural elements is essential for determining when machine autonomy should be narrowly constrained and when it can be more broadly applied. By leveraging these models, organizations can make informed decisions about the scope and limits of automation, ensuring that it aligns with operational goals and risk management strategies (As shown in Figure 2).

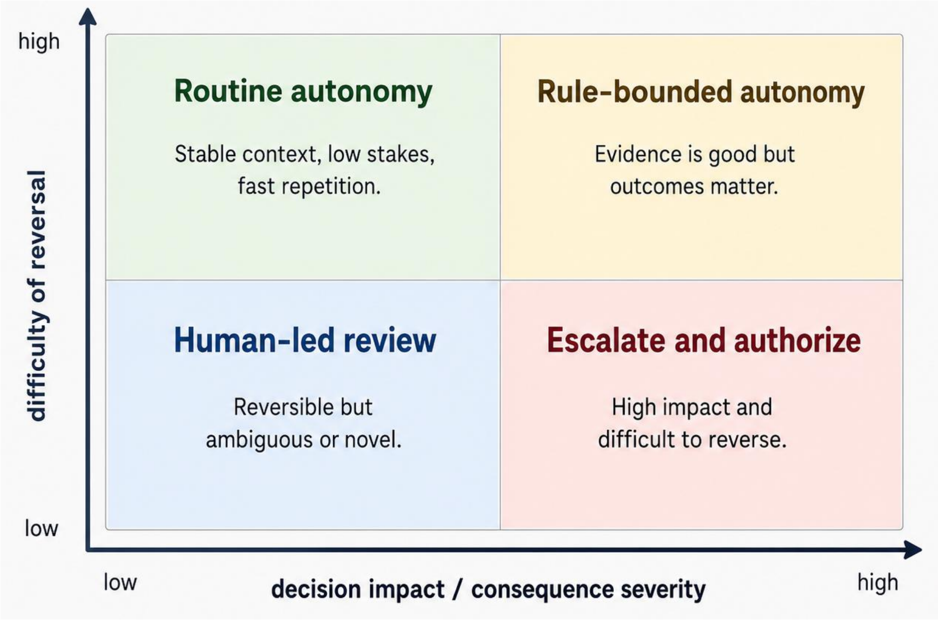

Figure 2. Autonomy Calibration Matrix.

4. Explanation, Audit and Accountable Use of Industrial AI

Explanation is often perceived as a supplementary communication tool, but in the context of industrial autonomy, it should be regarded as a fundamental control mechanism. The purpose of an explanation is to clarify the system's understanding of its state, the options it prioritized, the trade-offs that justified its recommendations, and the reasons why its actions adhered to predefined boundaries [12]. This does not necessitate exposing every mathematical detail; rather, it emphasizes the importance of making actions reviewable and contestable within a reasonable timeframe. By ensuring that explanations are accessible and actionable, organizations can better evaluate and refine their systems to align with operational goals and ethical standards. This approach fosters transparency and accountability, enabling stakeholders to assess the rationale behind decisions and intervene when necessary to prevent errors or inefficiencies.

Auditability serves as the institutional counterpart to explanation, providing a structured framework for tracing machine-supported decisions. Every decision made by an autonomous system should leave a comprehensive record that connects the observation basis, state interpretation, policy logic, authorization status, and the realized outcome [13, 14]. Such audit logs are essential for organizations to diagnose the root causes of poor performance, whether stemming from incorrect state assessments, flawed policies, rule conflicts, or execution failures. By linking accountability to continuous improvement rather than mere compliance, auditability ensures that systems evolve to meet operational demands effectively. This systematic approach to auditing not only enhances organizational learning but also strengthens the reliability and integrity of industrial AI applications.

The practical requirements for explanations vary significantly depending on the type of decision and the audience involved. For instance, a frontline operator may require concise explanations focused on controllable factors to make immediate adjustments. In contrast, a manager might need evidence demonstrating adherence to policy rules to ensure operational consistency. Meanwhile, governance teams often require detailed insights to compare actual outcomes with expected trade-offs, enabling them to determine whether existing rules need revision. These distinct explanation tasks highlight the inadequacy of a one-size-fits-all approach to explainability in Industry 4.0. Tailoring explanations to the specific needs of different stakeholders ensures that the information provided is both relevant and actionable, thereby enhancing decision-making processes across all levels of an organization.

Industrial studies underscore the importance of aligning technical reasoning with operational meaning, particularly when policies involve complex factors such as hidden degradation, resource constraints, or interconnected missions. In such cases, the most relevant explanations often pertain to the structure of the state and the associated resource trade-offs rather than isolated model coefficients. Well-designed explanations should bridge the gap between technical analysis and practical application, translating intricate computational processes into terms that are meaningful for operational decision-making. This alignment not only facilitates better understanding among stakeholders but also ensures that industrial AI systems are utilized effectively to achieve strategic objectives. By prioritizing clarity and relevance in explanations, organizations can optimize their use of autonomous systems while maintaining transparency and accountability.

5. Exception Handling and Escalation Design

If calibration defines routine authority, escalation design governs all scenarios that routine authority cannot effectively manage. Exceptions arise in situations where data are incomplete, recommendations conflict with localized knowledge, confidence levels drop below acceptable thresholds, objectives are at odds, or the operating environment changes too rapidly for the system to adapt. Many failures in industrial autonomy stem not from inaccuracies in prediction but from inadequately designed escalation pathways that fail to address these exceptions comprehensively. A robust escalation framework is essential to ensure that such challenges are managed effectively and do not compromise the overall system's functionality or safety.

An effective escalation design incorporates four critical properties to ensure its reliability and accountability. First, it must clearly define the triggers that initiate the escalation process, leaving no ambiguity about when intervention is required. Second, it must explicitly identify the actor or entity responsible for receiving and addressing the exception, ensuring accountability and clarity in the chain of command. Third, it must specify the information that must accompany the escalation, ensuring that the recipient has all the necessary context to make informed decisions. Finally, it must include mechanisms to update the system following resolution, preventing the recurrence of the same exception and enabling continuous improvement [15, 16]. While these properties may appear procedural, they are indispensable for maintaining accountability and ensuring that autonomous systems can operate effectively at scale without compromising safety or reliability.

Escalation mechanisms also serve as a safeguard against the unrealistic expectation of achieving full automation in complex environments. Smart factories, for instance, often feature a mix of heterogeneous assets, diverse objectives, and rare, unpredictable events. In such settings, a machine incapable of escalating issues gracefully may prove less effective than a simpler system that recognizes its limitations and seeks human intervention when necessary. This is particularly relevant in scenarios where policy learning is distributed across multiple interacting components or agents. The ability to escalate ensures that the system remains adaptable and responsive, even in the face of unforeseen challenges, thereby enhancing its overall utility and reliability.

The broader implication of escalation design is that human-in-the-loop control should not be relegated to a mere veto function at the end of the decision-making process [9]. Instead, human participation should be thoughtfully integrated through well-defined boundaries, triggers, and iterative learning loops. This integration ensures that autonomy enhances responsiveness while maintaining accountability. By designing systems that facilitate meaningful human involvement, organizations can strike a balance between leveraging the efficiency of automation and preserving the oversight necessary to address complex or unexpected scenarios effectively (As shown in Figure 3) (As shown in Table 2).

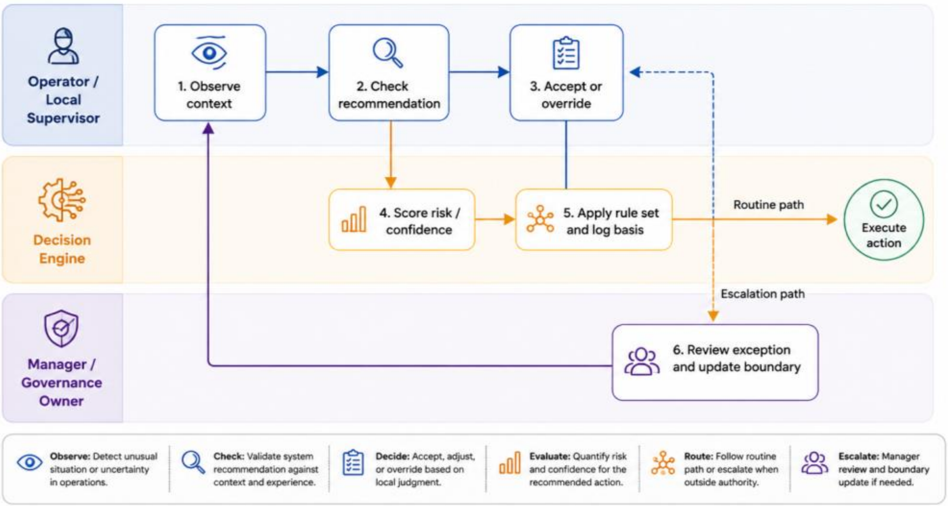

Figure 3. Exception-handling and Escalation Workflow.

Table 2. Common Governance Failure Modes and Practical Controls.

Failure mode	Operational symptom	Practical control
Opaque recommendation	Users ignore or bypass the system	Short explanation linked to state, option and trade-off
Automation overreach	Action executed beyond acceptable scope	Authority boundaries and rule-based caps
Silent drift	Past rules no longer fit current process	Periodic review and outcome monitoring
Exception overload	Managers receive too many escalations	Tiered thresholds and local triage
No learning from overrides	Same problem reappears repeatedly	Mandatory override logging and post-hoc review

6. Illustrative Application Domains

The framework is sufficiently versatile to address a wide range of industrial decision-making scenarios. In the domain of quality management, advanced models can be employed to interpret complex signals or images, enabling the identification of potential issues in production processes. These models can recommend specific interventions on production lines to enhance efficiency and quality, while ambiguous or uncertain cases are escalated to supervisors for further analysis. Similarly, in production control, scheduling engines can optimize localized decisions, such as task sequencing or resource allocation, while leaving more strategic decisions—such as adjustments to overtime, outsourcing, or prioritization of customer orders—to managerial discretion. In the energy sector, automated policies can dynamically adjust energy loads within predefined contractual boundaries, ensuring operational efficiency. However, when service levels are at risk of being compromised, these systems escalate the situation to human decision-makers for resolution, ensuring that critical thresholds are not breached [17].

Service and after-sales environments offer another compelling example of the framework's applicability. Self-service and field-service systems are often tasked with managing diverse types of customer demands, operating under conditions that may involve gradual deterioration or sudden, unexpected disruptions. This creates a complex decision-making challenge: determining which actions should be fully automated, which should be presented as recommendations for human approval, and under what circumstances scarce resources should be reallocated under supervisory oversight. For instance, in scenarios involving coupled inventory and operational policies, local decisions aimed at convenience or efficiency may sometimes conflict with broader, system-wide resource optimization goals. Addressing these conflicts requires a structured approach to decision rights, ensuring that both immediate and long-term objectives are balanced effectively.

Mission-oriented systems, particularly those operating under conditions of partial observability, introduce additional layers of complexity. In such systems, where certain states are hidden and outcomes evolve over time, decision-makers often require synthesized summaries, timely warnings, and clear escalation protocols rather than raw, unprocessed model outputs. For example, joint decisions involving mission objectives and intervention strategies highlight the critical role of authorization and adaptive learning. These systems must carefully balance short-term operational success against the potential for longer-term risks or exposures. This necessitates the development of robust frameworks that integrate real-time data analysis with strategic oversight, ensuring that decision-making processes remain both agile and accountable.

Learning-based systems, particularly those leveraging multi-agent reinforcement learning, underscore the importance of establishing explicit operational boundaries. While such systems can significantly enhance coordination and adaptability, organizations must clearly define the parameters within which agents are allowed to operate [18, 19]. This includes specifying the aspects of the system that agents may adapt, outlining mechanisms for resolving conflicts between agents, and determining the conditions under which policy updates require formal review and approval. In essence, the expansion of autonomy facilitated by learning-based systems is only sustainable if accompanied by a corresponding expansion in governance structures. This ensures that the benefits of increased autonomy do not come at the expense of accountability or operational stability.

7. Research Agenda

Decision-rights modeling represents a critical area of research within Industry 4.0, as current studies often prioritize architectural frameworks and algorithmic advancements over the nuanced allocation of authority. This imbalance highlights the need for a deeper exploration of governance structures and their adaptability across diverse operational environments. Effective decision-rights modeling requires the development of methodologies that can systematically compare governance designs, taking into account factors such as organizational hierarchy, resource distribution, and contextual variability. By addressing these gaps, researchers can contribute to more robust frameworks that optimize decision-making processes while ensuring accountability and operational efficiency.

Evidence-aware autonomy is an emerging research direction that challenges the assumption that improved predictive capabilities inherently justify increased automation. This perspective calls for a more comprehensive examination of how organizations calibrate autonomy based on the interplay of impact, reversibility, and evidence quality. For instance, high-impact decisions may necessitate stricter oversight and reduced autonomy, even when predictive accuracy is high, to mitigate risks associated with irreversibility [18]. Similarly, the quality of evidence supporting autonomous actions must be scrutinized to ensure reliability and alignment with organizational goals. By integrating these considerations, researchers can develop frameworks that balance automation with human oversight, fostering more adaptive and resilient systems.

Explanation under time pressure is a vital research area, particularly in industrial contexts where decisions often need to be made in real time or near real time. Traditional approaches to explanation prioritize theoretical completeness, which may not be feasible in high-pressure scenarios. Instead, operational efficiency must take precedence, ensuring that explanations are concise, actionable, and tailored to the immediate needs of decision-makers. This requires the development of tools and methodologies that can distill complex information into easily digestible formats without compromising critical insights.

By focusing on time-sensitive explanation mechanisms, researchers can enhance decision-making processes and support more effective responses to dynamic industrial challenges.

Lifecycle governance for learning systems addresses the limitations of static approval mechanisms in environments characterized by adaptive policies and evolving operational demands. Static governance frameworks often fail to accommodate the dynamic nature of learning systems, which require continuous updates to maintain alignment with organizational objectives. Research in this area should focus on identifying triggers for policy revisions, establishing protocols for override usage, and defining boundaries that can adapt over time. These efforts can lead to the creation of governance models that are both flexible and robust, ensuring that learning systems remain effective and accountable throughout their lifecycle [20].

Cross-level accountability is an essential research focus in the context of industrial autonomy, where decision-making spans multiple organizational levels, including operators, supervisors, managers, and model owners. Effective accountability frameworks must bridge the gap between local decision support systems and the broader organizational ownership of risk and performance [3, 10]. This involves developing mechanisms that clearly delineate responsibilities, facilitate communication across hierarchical levels, and ensure that risks are managed collaboratively. By addressing these challenges, researchers can contribute to the design of accountability systems that enhance transparency, foster trust, and optimize the integration of autonomous technologies within industrial operations.

8. Conclusion

This paper has argued that Industry 4.0 autonomy should be understood as a multifaceted design problem that extends beyond the narrow confines of technical capability. Specifically, the allocation of decision rights and the mechanisms for escalation play a pivotal role in determining the practical utility of autonomous systems. Autonomous operations are only as effective as the organizational frameworks that translate machine-generated interpretations into actionable, governable outcomes. This perspective underscores the importance of designing systems that not only perform tasks efficiently but also integrate seamlessly into broader governance structures, ensuring that automated decisions align with organizational objectives and ethical standards. By focusing on these dimensions, organizations can better harness the transformative potential of Industry 4.0 technologies while mitigating risks associated with unchecked automation.

Three foundational principles emerge as central to the effective deployment of autonomy within Industry 4.0 systems. First, the calibration of autonomy must consider the impact of decisions, their reversibility, and the quality of evidence supporting them. This ensures that autonomous systems operate within boundaries that reflect the significance of their actions and the reliability of their inputs. Second, explanation and auditability serve as critical mechanisms for accountability, enabling organizations to trace decision-making processes and validate outcomes. These features are essential for maintaining trust and transparency in automated systems. Third, the design of escalation pathways bridges the gap between scalable automation and meaningful human oversight. By embedding escalation protocols, organizations can ensure that human intervention is available when automated systems encounter scenarios that exceed predefined thresholds or require nuanced judgment. Collectively, these principles provide a robust framework for balancing efficiency, accountability, and control in autonomous operations.

Viewed through this lens, human-governed autonomy is not a compromise that diminishes the capabilities of automation but rather an essential organizational strategy that enhances its effectiveness. By integrating human oversight into the design of autonomous systems, organizations can achieve a dynamic balance that allows Industry 4.0 technologies to act swiftly, adapt continuously, and remain accountable in situations where consequences are significant. This approach fosters resilience and adaptability, enabling systems to learn from their environments while maintaining alignment with organizational goals. Furthermore, the emphasis on accountability and escalation ensures that autonomous systems remain responsive to ethical considerations and stakeholder expectations. Future research should explore the development of advanced escalation models, the integration of real-time audit mechanisms, and the application of these principles across diverse industries to further refine the interplay between automation and human governance.

References

1. J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.
2. K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.
3. M. P. Pacaux-Lemoine, D. Trentesaux, G. Z. Rey, and P. Millot, "Designing intelligent manufacturing systems through human-machine cooperation principles: a human-centered approach," *Computers & Industrial Engineering*, vol. 111, pp. 581–595, 2017.
4. L. Bainbridge, "Ironies of automation," in *Analysis, Design and Evaluation of Man–Machine Systems*, Pergamon, pp. 129–135, 1983.
5. F. Longo, L. Nicoletti, and A. Padovano, "Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context," *Computers & Industrial Engineering*, vol. 113, pp. 144–159, 2017.
6. S. Amershi et al., "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, May 2019.
7. D. Romero, P. Bernus, O. Noran, J. Stahre, and Å. Fast-Berglund, "The operator 4.0: Human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems," in *IFIP International Conference on Advances in Production Management Systems*, Cham: Springer International Publishing, pp. 677–686, Sep. 2016.
8. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
9. R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.
10. M. R. Endsley, "From here to autonomy: Lessons learned from human–automation research," *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017.
11. J. E. Teixeira and A. T. C. Tavares-Lehmann, "Industry 4.0 in the European Union: Policies and national strategies," *Technological Forecasting and Social Change*, vol. 180, p. 121664, 2022.
12. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
13. N. AI, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai, 2023.
14. B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.
15. S. A. Benraouane, *AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems*. Productivity Press, 2024.
16. I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, Jan. 2020.
17. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, Cambridge, vol. 1, no. 2, p. 25, 2018.
18. N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.
19. D. Acemoglu and P. Restrepo, "Automation and new tasks: How technology displaces and reinstates labor," *Journal of Economic Perspectives*, vol. 33, no. 2, pp. 3–30, 2019.
20. Y. Cheng, J. Wang, and Y. Wang, "A user-based bike rebalancing strategy for free-floating bike sharing systems: A bidding model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 154, p. 102438, 2021.