

Editorial: From copilot to autonomous agent: agentic AI in maintenance and the limits of delegation

Journal of Quality
in Maintenance
Engineering

821

1. The delegation moment

Consider a scenario that is closer than it appears. At 03:40, an anomaly detection model flags a bearing degradation signature on a critical compressor. An AI agent picks up the alert, queries the digital twin to estimate remaining useful life, generates a work order, checks inventory and finds the bearing out of stock, places a purchase requisition with the preferred supplier, reschedules two lower-priority jobs to free a qualified technician for the projected delivery window, and posts a summary to the planning board. The planner reads it over morning coffee. Everything worked.

The component capabilities required for this scenario already exist, and recent studies are beginning to integrate them into agent-like predictive-maintenance workflows, although evidence of robust end-to-end industrial autonomy remains limited (Di Maggio, 2025; Wang *et al.*, 2026). Commercial pressure to expand the scope of such autonomy is considerable and is likely to intensify. The question this editorial poses is therefore not whether agents can act but which of those actions ought to have required a human signature before execution, who in the organization decides that and on what basis. That is a question about delegation, and it admits a proposition that should guide everything that follows: decision authority may be delegated to an agent but organizational accountability cannot be.

This editorial is the fourth in a sequence. The first addressed representation and prediction, through digital twins and the gap between insight and prescription. The second addressed reasoning, asking whether generative AI in maintenance decision-making was hype, hazard, or genuine capability. The third addressed human capability, examining skills transfer in an era of AI-mediated work. The fourth addresses decision authority: what happens when the loop closes without a person inside it. This journal's readership, concerned with the quality and reliability of maintenance processes rather than merely their speed, is precisely the community that should define where the boundary of delegated authority belongs.

2. What “agentic” actually means in maintenance

Maintenance has automated decisions for decades. Condition-based triggers have long generated work orders, min-max inventory rules have reordered spare parts without human intervention and rule-based scheduling systems routinely resequence maintenance backlogs. A vendor demonstration that labels these established functions as “agentic” should therefore be assessed carefully. When the system merely executes predefined rules or workflows through a conversational interface, it represents conventional automation rather than a substantive advance in autonomous agency.

A genuinely agentic system is distinguished by a combination of characteristics that classical automation lacks. Classical agent theory emphasizes autonomy, reactivity, proactiveness and social ability (Wooldridge and Jennings, 1995). Contemporary large language model (LLM)-based agent architectures extend these characteristics through goal-directed planning, autonomous selection and sequencing of actions, access to multiple tools and enterprise systems, persistence of state and memory across tasks and adaptation in response to feedback (Di Maggio, 2025). In maintenance, a single decision episode may therefore span the computerized maintenance management system (CMMS), enterprise resource planning (ERP), digital twin and procurement systems. Such action must occur within a defined authority boundary. An agent without an explicit authority boundary is not more capable; it is merely ungoverned. The requirements literature on digital twin-driven



Journal of Quality in Maintenance
Engineering
Vol. 32 No. 4, 2026
pp. 821-826
© Emerald Publishing Limited
e-ISSN: 1758-7832
p-ISSN: 1355-2511
DOI 10.1108/JQME-10-2026-166

autonomous maintenance similarly treats autonomy as a system-level architectural capability rather than a feature of any single model (Khan *et al.*, 2020). Determining the authority within which that capability may operate is therefore also an organizational governance task.

A separate issue, better treated as an audit and oversight risk than as a defining feature of agency, is the ability of contemporary agents to generate fluent natural-language explanations. Such explanations may support communication and review, but they should not be treated as faithful evidence of the agent's reasoning or as proof that its decision is correct. A rationale can remain coherent and persuasive even when the underlying data, assumptions, tool responses, or planning steps are flawed. By contrast, you can usually inspect the logic and parameters of a conventional min-max rule directly. Agentic systems therefore introduce a distinctive oversight risk: they can make erroneous decisions appear well reasoned. The hype-hazard-capability triage proposed in the second editorial remains applicable, but the stakes are higher because a flawed premise may now be translated into actions across interconnected systems rather than remain confined to a report or recommendation.

3. A delegation taxonomy

If delegation is the right question, it needs structure. We propose classifying maintenance decisions along two axes that jointly determine the degree of autonomy that can be defensibly granted.

The first axis is decision risk exposure, which combines the potential consequence of the decision with its reversibility. Re-sorting a work order backlog is generally low consequence and relatively easy to reverse, although a poorly sequenced backlog can still delay safety-critical or production-critical work. Deferring a preventive task consumes safety margin in ways that may surface only months later, and adjusting operating parameters on a safety-critical asset may be irreversible in the only sense that matters. The second axis is independent verifiability: whether the decision can be checked by a human reliably and at a cost meaningfully lower than making it. A proposed schedule can be scanned in seconds against a planner's mental model. A remaining-useful-life estimate synthesized from multiple data sources may be costly to verify independently, requiring model validation, diagnostic corroboration, or reproduction of parts of the analysis.

Crossing these two axes yields the four decision-authority modes presented in Figure 1, in the tradition of graduated levels of automation (Parasuraman *et al.*, 2000). This framework converts a general debate about AI autonomy into a maintenance decision-design problem, and we offer it as the editorial's central contribution.

Decisions of intermediate risk are boundary cases rather than a fifth category: their treatment depends on asset criticality, authorization limits and the organization's risk tolerance, which is precisely why authority assignment must remain an explicit design act rather than a default inherited from the technology.

Two observations give the taxonomy practical significance. First, assigning a decision class to an authority mode is an organizational design choice rather than an inherent property of the agent. The same technical capability may be permitted to operate autonomously in one plant but limited to generating recommendations in another because asset criticality, regulatory requirements, review capacity and organizational risk tolerance differ. External certification of technical performance may inform this choice, but responsibility for the operational authority granted must remain with the asset owner. Second, consequential failures may result not only from technical malfunction but also from authority misassignment. In such cases, the agent may perform its assigned task as designed, while the organization has placed that task in an authority mode that permits more autonomy than the associated risk and verifiability can justify.

4. Assuring the agent: the quality gap

Here the topic becomes unavoidably a quality-in-maintenance problem. The instruments this community has developed over decades, including reliability-centered maintenance (RCM)

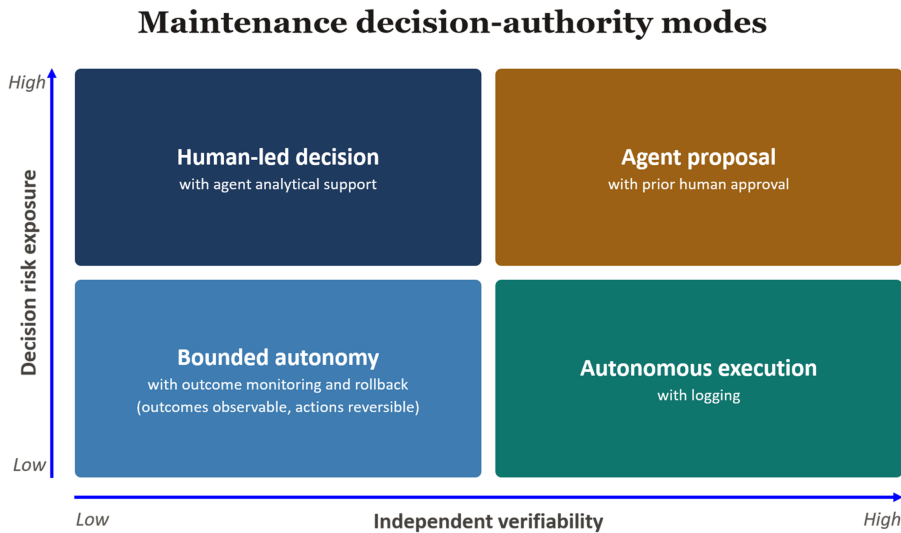


Figure 1. Maintenance decision-authority modes based on risk exposure and independent verifiability

logic, failure mode and effects analysis (FMEA), work-order audit trails and asset management systems aligned with International Organization for Standardization (ISO) 55000 and ISO 55001 (ISO, 2024a, b), are built around clearly defined organizational roles, personnel competence, documented processes and auditable management systems. These foundations remain essential, but they do not by themselves provide a complete assurance regime for autonomous agents whose decisions may be generated and executed across interconnected systems. Emerging AI management and risk frameworks are beginning to address this gap, although their translation into maintenance-specific authorization, validation and oversight practices remains incomplete (NIST, 2023; ISO/IEC, 2023).

Three gaps follow. The first concerns evidence. Assurance should rest not on the agent's generated self-explanation, which may not faithfully represent its internal process, but on decision provenance: a durable record comprising the system and model version, the input data and timestamps, the records and evidence retrieved, the policy and constraint set in force, the tool calls made and the responses received, the recommendation generated, the action executed, any human approvals or overrides and the subsequent outcome. Incident investigation protocols must be able to reconstruct this chain and distinguish a corrupted input from a flawed plan from a failed tool response.

The second gap concerns scale. Human decisions can also propagate rapidly through automated systems, but an autonomous agent adds repetition: one bad premise, such as a corrupted sensor feed or a stale bill of materials, can be executed across every asset within the agent's scope before the first review cycle begins. Quality assurance designed around weekly planning meetings cannot govern decisions made in milliseconds. Blast-radius constraints, rate limits and automatic circuit breakers are not exotic safeguards; they are the agentic equivalent of lockout-tagout.

The third gap concerns responsibility. When an agent-initiated action goes wrong, the causal chain runs through the model vendor, the integrator, the engineer who configured the tools, the manager who assigned the authority mode and the technician who executed the output and each will point at the others. Maintenance organizations have addressed analogous problems through qualification regimes for personnel and functional-safety regimes for systems (IEC, 2010). By analogy, we propose that the corresponding instrument for agents is an operational authorization built on a validated competence envelope: an explicit statement of what the agent is authorized to

decide, the conditions under which that authorization holds, and the evidence supporting it. A material change in the model, its data, its prompts, its tool configuration, or its system integration should trigger revalidation before the authorization is renewed.

5. Supervision and the deskilling paradox

The standard reassurance offered for all of this is “human in the loop.” The phrase deserves more skepticism than it receives. Decades of human factors research document what happens to human oversight of highly reliable automation: vigilance and situation awareness can deteriorate, reliance can become miscalibrated and operators may over-rely on highly reliable automation or struggle to intervene effectively when required (Parasuraman and Riley, 1997; Lee and See, 2004; Endsley, 2017). A propose-and-approve mode in which the human endorses nearly every proposal without genuine scrutiny is an autonomous mode with extra steps and a comforting audit trail. If oversight is to constitute meaningful human control rather than nominal presence (Santoni de Sio and van den Hoven, 2018), it must be designed: sampling regimes that keep reviewers genuinely engaged, interfaces that surface the agent’s uncertainty rather than its confidence and workloads that leave time for scrutiny rather than throughput alone.

The deeper problem compounds over time and was anticipated in Bainbridge’s (1983) ironies of automation: the judgment required to supervise an automated system is built through the very routine practice the system removes. The previous editorial examined how skills are transferred when AI mediates the work; agentic AI sharpens that concern into a paradox. The planner who can sense that a proposed deferral is subtly wrong acquired that sense by making deferral decisions for years. Delegate the routine work and the pipeline that produces qualified supervisors runs dry. An organization may gradually consume its inherited expertise before recognizing that few employees remain capable of meaningfully challenging the agent, and by then the authority modes are no longer a choice.

The design goal, for both problems, is calibrated reliance: organizations should engineer for appropriate trust, not maximal trust (Lee and See, 2004). That means deliberately retaining human practice on a subset of decisions the agent could handle, treating disagreement between agent and human as a learning signal to be investigated rather than a friction to be optimized away, and managing supervisory competence as an asset with a depreciation schedule.

6. A research agenda

The questions above are researchable, and this community is well equipped to research them. We group the priorities under three headings.

Technical assurance. First, verification of agent plans and actions: methods for formally checking agent-generated plans against RCM and FMEA logic, maintenance strategy constraints and safe operating envelopes, so that assurance rests on the checked plan and its provenance rather than on the narrative explanation. Second, FMEA of the decision layer itself: enumerating the failure modes of goal-directed planners, including premise corruption, objective misspecification and tool-response errors, together with their detection mechanisms. Third, agent performance measurement: defining the maintenance key performance indicators (KPIs) of an agent and benchmarking against human planners in ways that resist gaming, counting the review burden the agent generates as well as the labor it saves.

Human oversight. Fourth, the human factors of supervision in maintenance settings: interface design, approval fatigue, sampling strategies and the conditions under which oversight of mostly-correct systems remains genuine rather than nominal. Fifth, competence retention under delegation: how organizations preserve the supervisory judgment that agentic systems presuppose, and how that retention is measured.

Governance. Sixth, empirically grounded studies of delegation design: which decisions real maintenance organizations delegate, which they reclaim after incidents and what organizational factors predict these differences. The field now needs empirical studies that test,

refine and establish the boundary conditions of conceptual classifications such as the one proposed here. Seventh, the economics of delegation: models trading the cost of review against the cost of error across asset classes, locating the optimal authority boundary as a function of risk exposure, verifiability and review capacity. Eighth, authorization regimes: what an operational authorization framework for agents would contain, including competence envelopes, revalidation triggers and withdrawal criteria, developed in dialogue with asset management and AI management standards (ISO, 2024a; ISO/IEC, 2023).

7. Autonomy can be delegated; accountability cannot

The title marks a substantive transition. A copilot recommends, while the human or organization that authorizes and acts on the recommendation remains answerable for the outcome. An autonomous agent can initiate and execute actions yet prevailing ethical and legal frameworks do not generally treat the AI system itself as a bearer of professional, moral or organizational accountability. Responsibility instead remains distributed among the human and institutional actors who design, provide, deploy, authorize and oversee the system, even though adaptive and autonomous behavior can complicate conventional responsibility attribution (Matthias, 2004; Santoni de Sio and van den Hoven, 2018; European Union, 2024; Council of Europe, 2024). The transition from copilot to autonomous agent is therefore not achieved merely when technical performance becomes adequate. It becomes defensible when the organization establishes an explicit authority boundary, retains meaningful human control, preserves the provenance of decisions and actions, validates the agent's defined competence envelope and specifies triggers for intervention, suspension and reauthorization (NIST, 2023; ISO/IEC, 2023). Under these conditions, decision authority may be delegated, but accountability remains with the people and organizations that grant, govern and retain the power to withdraw that authority.

The maintenance engineering community should feel a certain confidence here. Structured failure analysis, reliability thinking, qualification regimes and disciplined incident investigation are its native tools, refined over half a century. The governance of autonomous decision-making is not a problem this field must import from elsewhere; it is a problem this field is unusually well placed to lead. The alternative, inheriting delegation frameworks designed by vendors under commercial pressure, is how authority misassignment becomes an industry default rather than an avoidable error.

Agents are coming to maintenance, and much of what they will do is genuinely valuable. The task now is to decide, deliberately and on our own professional terms, what they may be trusted to do alone, under what safeguards and with what evidence. Autonomy can be delegated. Accountability cannot and the organizations that internalize that asymmetry first will be the ones that capture the value of agentic AI without inheriting its failures.

Mohamed Ben-Daya

Acknowledgments

LLMs were used for English editing and searching for suitable references.

References

- Bainbridge, L. (1983), "Ironies of automation", *Automatica*, Vol. 19 No. 6, pp. 775-779, doi: [10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8).
- Council of Europe (2024), *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, Council of Europe Treaty Series No. 225, Council of Europe, Strasbourg.

- Di Maggio, L.G. (2025), "Toward autonomous LLM-based AI agents for predictive maintenance: state of the art, challenges, and future perspectives", *Applied Sciences*, Vol. 15 No. 21, 11515, doi: [10.3390/app152111515](https://doi.org/10.3390/app152111515).
- Endsley, M.R. (2017), "From here to autonomy: lessons learned from human-automation research", *Human Factors*, Vol. 59 No. 1, pp. 5-27, doi: [10.1177/0018720816681350](https://doi.org/10.1177/0018720816681350).
- European Union (2024), "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence", *Official Journal of the European Union*, L 2024/1689.
- IEC (2010), *IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*, International Electrotechnical Commission, Geneva.
- ISO (2024a), *ISO 55000: Asset Management – Vocabulary, Overview and Principles*, International Organization for Standardization, Geneva.
- ISO (2024b), *ISO 55001: Asset Management – Asset Management System – Requirements*, International Organization for Standardization, Geneva.
- ISO/IEC (2023), *ISO/IEC 42001: Information Technology – Artificial Intelligence – Management System*, International Organization for Standardization, Geneva.
- Khan, S., Farnsworth, M., McWilliam, R. and Erkoyuncu, J. (2020), "On the requirements of digital twin-driven autonomous maintenance", *Annual Reviews in Control*, Vol. 50, pp. 13-28, doi: [10.1016/j.arcontrol.2020.08.003](https://doi.org/10.1016/j.arcontrol.2020.08.003).
- Lee, J.D. and See, K.A. (2004), "Trust in automation: designing for appropriate reliance", *Human Factors*, Vol. 46 No. 1, pp. 50-80, doi: [10.1518/hfes.46.1.50.30392](https://doi.org/10.1518/hfes.46.1.50.30392).
- Matthias, A. (2004), "The responsibility gap: ascribing responsibility for the actions of learning automata", *Ethics and Information Technology*, Vol. 6 No. 3, pp. 175-183, doi: [10.1007/s10676-004-3422-1](https://doi.org/10.1007/s10676-004-3422-1).
- NIST (2023), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, National Institute of Standards and Technology, Gaithersburg, MD.
- Parasuraman, R. and Riley, V. (1997), "Humans and automation: use, misuse, disuse, abuse", *Human Factors*, Vol. 39 No. 2, pp. 230-253, doi: [10.1518/001872097778543886](https://doi.org/10.1518/001872097778543886).
- Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000), "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, doi: [10.1109/3468.844354](https://doi.org/10.1109/3468.844354).
- Santoni de Sio, F. and van den Hoven, J. (2018), "Meaningful human control over autonomous systems: a philosophical account", *Frontiers in Robotics and AI*, Vol. 5, 15, doi: [10.3389/frobt.2018.00015](https://doi.org/10.3389/frobt.2018.00015).
- Wang, C., Wu, T. and Bao, J. (2026), "Agentic digital twin-embedded maintenance methodology for energy equipment: a self-evolving operational paradigm", *Journal of Manufacturing Systems*, Vol. 84, pp. 100-116, doi: [10.1016/j.jmsy.2025.11.020](https://doi.org/10.1016/j.jmsy.2025.11.020).
- Wooldridge, M. and Jennings, N.R. (1995), "Intelligent agents: theory and practice", *The Knowledge Engineering Review*, Vol. 10 No. 2, pp. 115-152, doi: [10.1017/s02698889000008122](https://doi.org/10.1017/s02698889000008122).