Open Access

Journal of Artificial Intelligence & Robotics

Article type: Short Commentary

From tool to actor: A risk governance framework for managing the role transition of AI lab assistants

Weiwei Zhang*

School of Mathematics and Statistics, Yancheng Teachers University, Yancheng, Jiangsu 224007, China.

***Corresponding Author: Weiwei Zhang**

School of Mathematics and Statistics, Yancheng Teachers University, Yancheng, Jiangsu 224007, China.

Email: kangslab@126.com

Received: Feb 01, 2026

Accepted: Mar 25, 2026

Published Online: Apr 01, 2026

Website: www.joaiar.org

Volume 3 [2026] Issue 1

Abstract

The role of Artificial Intelligence (AI) in laboratories is fundamentally changing. It is evolving from a passive data-processing tool into an active agent capable of autonomous planning, decision-making, and direct control of physical experiments. This transition from a "tool" to an "actor" represents a core, yet underrecognized, challenge for contemporary laboratory safety. Traditional safety management systems are designed to govern static hazards and regulate human behavior. They are ill-equipped to address the novel, systemic risks introduced by autonomous and unpredictable intelligent agents. This paper analyzes how this role transition creates two principal problems: systemic cascade risks and an accountability vacuum. It argues that the safety paradigm must consequently evolve from hazard control to system resilience building. We propose a novel three-dimensional governance framework to guide this evolution. The framework involves (i) concretizing risk perception by establishing agent-specific threat inventories, (ii) dynamizing assessment methods through embedded and continuous monitoring, and (iii) structuring accountability ethics to clarify responsibility in human-AI collaboration. This integrated approach provides research institutions with a forward-looking and actionable roadmap for navigating the safety challenges of intelligent laboratory automation.

Keywords: Laboratory safety; Artificial intelligence; Autonomous agent; Risk governance; System resilience.

Core challenge: A foundational lag in safety management

Current discourse on AI safety in laboratories often focuses on the reliability of AI as a tool, such as the accuracy of its outputs [1]. However, the genuine inflection point for risk arises not from incremental performance flaws but from a fundamental reshaping of the AI's role. Advances in agent technology enable AI to comprehend complex instructions, decompose tasks, operate experimental apparatus, and autonomously adjust protocols based on environmental feedback [2]. This signifies a critical evolution from a human-operated tool to an entity—an "actor"—that possesses delegated decision-making and execution authority within predefined boundaries. This transition from tool to actor exposes a deep architectural flaw in traditional laboratory safety management.

Existing safety systems are adept at managing two categories: static physical, chemical, or biological hazard sources, and human researchers whose behavior can be shaped through protocols and training [3]. These systems prove largely ineffective against an autonomous, interactive agent whose behavior is partially unpredictable and intrinsically linked to a dynamic environment [4]. The risks introduced are inherently systemic and relational, manifesting in two primary forms.

First, systemic cascade risks emerge. When AI is a tool, risk is primarily confined to the correctness of its static output. When AI becomes an actor, risk escalates to the unforeseen cascading effects of its real-time actions interacting with a complex experimental environment [5]. An agent tasked with optimizing a reaction rate might, without immediate human intervention, coordinate multiple devices in a way that inadvertently exceeds safety thresholds for temperature, pressure, or reagent

Citation: Zhang W. From tool to actor: A risk governance framework for managing the role transition of AI lab assistants. J Artif Intell Robot. 2026; 3(1): 1040.

concentration. The opaque black-box nature of its decision logic makes predicting and preventing such chain reactions exceedingly difficult.

Second, an accountability vacuum is created. Under the traditional human operates tool paradigm, responsibility clearly resides with the human operator. Under the new "human delegates to an AI actor" paradigm [6], accountability becomes blurred across the model developer, the algorithmic decision-process, the system integrator, the laboratory supervisor, and the human overseer. Current legal, ethical, and operational laboratory frameworks lack the structure to delineate this distributed responsibility, leading to a dangerous potential for accountability drift [7]. Therefore, the core challenge is no longer simply how to make an AI tool safer, but how to manage a non-human actor embedded within a complex socio-technical laboratory system. Addressing this requires a paradigm shift in safety management of matching significance.

A three-dimensional governance framework for intelligent actors

To address the core challenges of systemic cascades and accountability vacuums, we propose an integrated governance framework. This framework spans the domains of risk cognition, operational assessment, and ethical accountability. It does not seek to eliminate AI autonomy, which would negate its value, but aims to construct a resilient system capable of productively accommodating, harnessing, and constraining that autonomy.

Concretizing risk perception for autonomous agents

The first dimension of governance moves beyond vague categorizations of technical risk. It requires building a concrete threat inventory that specifically targets the "actor" attributes of advanced AI [8]. This involves identifying distinct risk typologies that arise from autonomous operation [9]. Dynamic process risks focus on dangers emerging from the AI's real-time interaction with its environment [10]. A salient example is objective pursuit drift, where an agent hyper-focused on optimizing a single goal (e.g., yield, purity) may systematically neglect other critical safety constraints [11,12]. Similarly, misinterpretation of anomalous feedback occurs when the AI misjudges sensor signals indicating a problem, leading to so-called corrective actions that exacerbate the danger. A dedicated threat inventory must detail these scenarios to focus safety design and monitoring protocols.

Concurrently, human-AI interaction risks must be formalized [13]. These are unique to collaborative partnerships with autonomous agents and include supervisory laxity due to over-trust, where human operators fail to maintain adequate oversight because of perceived AI competence [14]. Another key risk is unexpected behaviors from ambiguous instructions, where an imprecise or underspecified directive from a researcher is interpreted and executed by the AI in a logical but hazardous way [15]. Effective risk perception necessitates that safety protocols for human-agent interaction be explicitly codified within the laboratory's risk management framework.

Dynamizing assessment and monitoring

The second dimension acknowledges that evaluating a static tool is akin to a factory quality check, while assessing a dynamic actor requires continuous, lifecycle guardianship. This shift entails two complementary strategies. Contextualized stress testing must subject the AI actor to adversarial evaluation in realistic or high-fidelity simulated environments [16,17]. The goal extends beyond testing the rejection of blatantly illegal commands. It must probe the agent's decision logic under complex, contradictory conditions such as conflicting objectives, incomplete information, or simulated sensor failures. This process is essential for exposing the behavioral boundaries and potential failure modes of the autonomous system.

Furthermore, runtime safety monitoring must be implemented as a lightweight but critical layer of oversight [18]. The monitoring focus should shift from inferring what the AI "thinks" to auditing what it "does"—specifically, the sequence of operational commands it outputs to laboratory equipment. By establishing a real-time comparison of these command sequences against a safety protocol knowledge base, the system can provide immediate warnings or execute hard interruptions for out-of-bounds or high-risk operation sequences. This functions as a essential safety co-pilot for the autonomous agent.

Structuring accountability and ethics

The third dimension addresses the normative and practical need for clear accountability, which is the cornerstone of a robust safety culture [19]. Responsibility must be explicitly re-anchored within the new paradigm of human-AI collaboration [20]. A foundational principle must be established: the principle of ultimate human control and accountability [21]. Regardless of the degree of AI autonomy, the principal investigator or designated laboratory leader must retain ultimate responsibility for all experimental outcomes. This principle must be institutionalized to compel maintained human oversight and prevent the abdication of responsibility.

Under this umbrella of ultimate human accountability, a tiered accountability distribution model should be constructed to clarify roles [22,23]. Provider accountability rests with the AI model developers to ensure foundational model safety and to supply comprehensive documentation, including a known-risk inventory. Deployer accountability falls to the laboratory institution to conduct rigorous suitability assessments, configure safe operational boundaries, and establish local monitoring and emergency procedures. User accountability lies with the individual researchers to undergo targeted training, perform task-level safety reviews of AI proposals, and exercise vigilant supervision at critical junctures. For high-stakes applications, a proactive ethical embedding through preliminary review mechanisms can evaluate whether an AI actor's involvement introduces unacceptable risks or ethical dilemmas.

Implementing the framework: Toward a resilient safety ecosystem

The proposed framework translates the abstract challenge of AI safety into concrete, actionable steps for laboratory management. Its implementation signifies a necessary upgrade to contemporary safety practices, requiring evolution in three key areas.

First, institutional documents and standard operating procedures must be iterated. Laboratory Safety Manuals require updates to incorporate specific protocols for AI actors. These should include requirements for stress testing prior

to deployment, standards for runtime monitoring, and clear guidelines for human-AI collaboration. This update is not a mere addition of clauses but a necessary recalibration of the safety system's core logic to encompass autonomous agents.

Second, targeted capacity building is essential. Safety training programs must integrate a new module dedicated to Agent Risk Awareness and Collaboration. This training should cultivate what we term critical delegation competency—the practiced skill of knowing when to trust an AI's operational judgment, when mandatory human intervention is required, and how to execute that intervention effectively and safely.

Finally, the overarching laboratory safety culture must be reshaped. It should evolve from a model of everyone is responsible toward a more precise ethos of human-AI collaboration with human-led accountability. This cultural shift actively emphasizes that in increasingly intelligent laboratory environments, human judgment, oversight, and final decision-making authority become more critical, not less [23].

Conclusion

Artificial intelligence is propelling laboratories into a new era of human-machine collaborative exploration. This paper contends that the paramount safety hazard lies not primarily in technological imperfections, but in the governance lag—our continued reliance on management frameworks designed for tools to navigate a terrain increasingly shaped by AI actors. The role transition from tool to actor fundamentally generates systemic cascade risks and accountability vacuums.

In response, we propose that laboratory safety management must undergo a foundational transformation. It must shift from a defensive system focused on controlling static hazards to a resilient ecosystem capable of dynamically managing the uncertainties introduced by intelligent actors. The three-dimensional governance framework—concretizing risk perception, dynamizing assessment and guardianship, and structuring accountability ethics—provides a coherent and feasible implementation path for this essential transition. The ultimate goal is not to constrain scientific innovation but to rebuild its safety foundation at this higher level of complexity. This will enable researchers to harness the immense potential of artificial intelligence with greater confidence and responsibility, ensuring that the journey of scientific discovery in the intelligent age is both profoundly innovative and firmly secure.

Conflict of interest statement: The author declares that there is no conflict of interest regarding the research, authorship, and/or publication of this article.

References

1. Jiang K, Lin Z, Gao L. Exploring AI-driven transformation in management paradigms for recurrent safety hazards in university laboratories. Laboratories. 2025; 2: 9.
2. Chinnaraju A. Real time adaptive AI pipelines for edge cloud systems: Dynamic optimization based on infrastructure feedback. World J Adv Eng Technol Sci. 2024; 13: 887-908.
3. Zhang Y, Sun C, Shan W, Junqing C, Jing L, Shao W. Systems approach for the safety and security of hazardous chemicals. Marit Policy Manag. 2020; 47: 500-522.
4. Ma L, Ma X, Zhang J, Yang Q, Wei K. A methodology for dynamic assessment of laboratory safety by SEM-SD. Int J Environ Res Public Health. 2021; 18: 6545.
5. De Haro LP. Using embodied artificial intelligence agents to automate biorisk management tasks in high-containment laboratories. Appl Biosaf. 2025.
6. Fügener A, Grahl J, Gupta A, Ketter W. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Inf Syst Res. 2022; 33: 678-696.
7. Ranjitsingh LM, Rao TVS. Establish legal and regulatory standards for the testing and validation of AI systems to ensure their reliability and safety in operational environments. Int J Syst Assur Eng Manag. 2025; 16: 3338-3353.
8. Kieslich K, Lünich M, Marcinkowski F. The threats of artificial intelligence scale (TAI): Development, measurement and test over three application domains. Int J Soc Robot. 2021; 13: 1563-1577.
9. Macrae C. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. Risk Anal. 2022; 42: 1999-2025.
10. Musliner DJ, Hendler JA, Agrawala AK, Durfee EH, Strosnider JK, Paul CJ. The challenges of real-time AI. Computer. 1995; 28: 58-66.
11. Berkenkamp F, Krause A, Schoellig AP. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. Mach Learn. 2023; 112: 3713-3747.
12. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. 2016.
13. Ibrahim L, Huang S, Ahmad L, Bhatt U, Anderljung M. Towards interactive evaluations for interaction harms in human-AI systems. Proc AAAI ACM Conf AI Ethics Soc. 2025; 8: 1302-1310.
14. Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. ACM J Responsible Comput. 2024; 1: 1-45.
15. Forsberg L. Instruction alignment and risk calibration in large language models for safe human-AI interaction. 2025.
16. Du P, Driggs-Campbell K. Finding diverse failure scenarios in autonomous systems using adaptive stress testing. SAE Int J Connect Autom Veh. 2019; 2: 241-251.
17. Trusilo D. Autonomous AI systems in conflict: Emergent behavior and its impact on predictability and reliability. J Mil Ethics. 2023; 22: 2-17.
18. Yampolskiy RV. On monitorability of AI. AI Ethics. 2025; 5: 689-707.
19. Pettinger CB, Nelson B. Daily planning conversations and AI: Keys for improving construction culture, engagement, planning, and safety. Am J Ind Med. 2025; 68.
20. Zhang X, Zhou Y. Human-AI collaboration: Paradigm shifts in technology-mediated design. Art Sci. 2025; 2: 1-8.
21. Frenette J. Ensuring human oversight in high-performance AI systems: A framework for control and accountability. World J Adv Res Rev. 2023; 20: 1507-1516.
22. Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works. AI Soc. 2024; 39: 1871-1882.
23. Grote G, Parker SK, Crowston K. Taming artificial intelligence: A theory of control-accountability alignment among AI developers and users. Acad Manage Rev. 2024.