Journal of Information Systems Engineering and Management
2026, 11(1s) e-ISSN: 2468-4376 https://www.jisem-journal.com/ **Research Article**

AI-Assisted Virtual Machine Right-Sizing with Human Oversight: A Comprehensive Technical Analysis

Priyadarshni Shanmugavadivelu

Birla Institute of Technology and Science, Pilani, India

ARTICLE INFO

Received: 23 Jan 2026

Revised: 28 Jan 2026

ABSTRACT

The challenge of virtual machine right-sizing remains significant because cloud computing environments are characterized by unpredictable workloads, sensitivity to performance, and massive deployments that impact resource optimization strategies. Manual right-sizing relies on periodic reviews, fixed utilization limits, and human intuition and is therefore slow, inconsistent, and difficult to apply to large enterprise infrastructures. On the other hand, completely automated right-sizing systems are operationally unsafe as well because hyper-aggressive resizing behavior may inadvertently lead to performance impairment, service level agreement breach, or even production workload instability. Right-sizing systems supported by AI deal with these limitations by operating on workload telemetry, historical use trends, and performance indicators in real time to produce evidence-based suggestions for optimal VM configurations. These systems do not implement change independently. Instead, they provide recommendations along with confidence scores and impact ratings, and are able to process high-volume data using artificial intelligence while avoiding the risks associated with fully automated operations in sensitive infrastructure environments. Human-in-the-loop governance models position AI as an augmentation of human decision-making rather than a replacement. Cloud platform operators retain veto authority, exercise discretion, and introduce customer-specific factors when evaluating recommendations. This cooperative approach enables improved cost efficiency and more effective resource utilization without sacrificing reliability or operational confidence in cloud platform management.

Keywords: Virtual Machine Right-Sizing, Human-In-The-Loop Governance, AI-Driven Workload Optimization, Cloud Infrastructure Management, Explainable Artificial Intelligence

1. Introduction

The rapid growth of cloud computing infrastructure has introduced significant challenges in optimizing virtual machine resources within enterprise data centers. As organizations increasingly rely on cloud environments to run critical workloads, effective allocation of computational resources is essential for maintaining operational efficiency and controlling costs. A study published in *Future Generation Computer Systems* highlights that cloud data centers consume substantial amounts of electrical energy, much of which is driven by inefficient resource utilization and the overhead of operating idle or underutilized virtual machines [1]. This energy consumption represents not only a direct financial cost but also a growing environmental burden, making resource optimization both an economic and sustainability imperative. The complexity of aligning VM configurations with real workload demands across diverse applications has historically pushed organizations toward either manual approaches that do not scale or fully automated systems that introduce unacceptable operational risk. Recent advances in artificial intelligence and machine learning offer new opportunities to address this challenge through intelligent recommendation systems. The *Guidelines for Human-AI Interaction* project by Microsoft Research emphasizes that effective AI systems must support human oversight to ensure transparency

and appropriate levels of automation [2]. This principle is particularly relevant to infrastructure management, where incorrect decisions can rapidly propagate through production environments. This paper examines the technical foundations, system design, and governance models required to implement AI-assisted VM right-sizing while balancing automation with human accountability and operational responsibility.

2. Limitations of Manual Virtual Machine Right-Sizing Approaches

Traditional approaches to VM right-sizing in enterprise environments suffer from fundamental limitations that become increasingly pronounced as deployment scale grows. Manual right-sizing typically involves infrastructure engineers periodically reviewing VM utilization metrics, comparing observed values against predefined thresholds, and making resizing decisions based on professional judgment and experience. Research on predictive analytics for resource management in IaaS clouds shows that conventional capacity planning frameworks fail to capture the dynamic nature of cloud workloads, where resource demands fluctuate unpredictably due to user behavior, seasonal effects, and application lifecycle events [3]. As a result, manual reviews are inherently reactive, identifying misallocation only after inefficiencies have persisted long enough to generate unnecessary cost. A further limitation of manual approaches is their reliance on fixed utilization thresholds. These predefined criteria are insufficiently responsive to temporal workload variation, burst patterns, and the complex relationship between resource consumption and application performance. A comprehensive survey published in the *Journal of Internet Services and Applications* identifies resource management as a central research challenge in cloud computing, emphasizing the need to account for workload characteristics, performance requirements, and interdependencies across multiple resource dimensions, including compute, memory, storage, and network [4]. Such multidimensional relationships cannot be adequately captured using static, threshold-based models, leading to decisions that either leave resources underutilized or introduce performance bottlenecks.

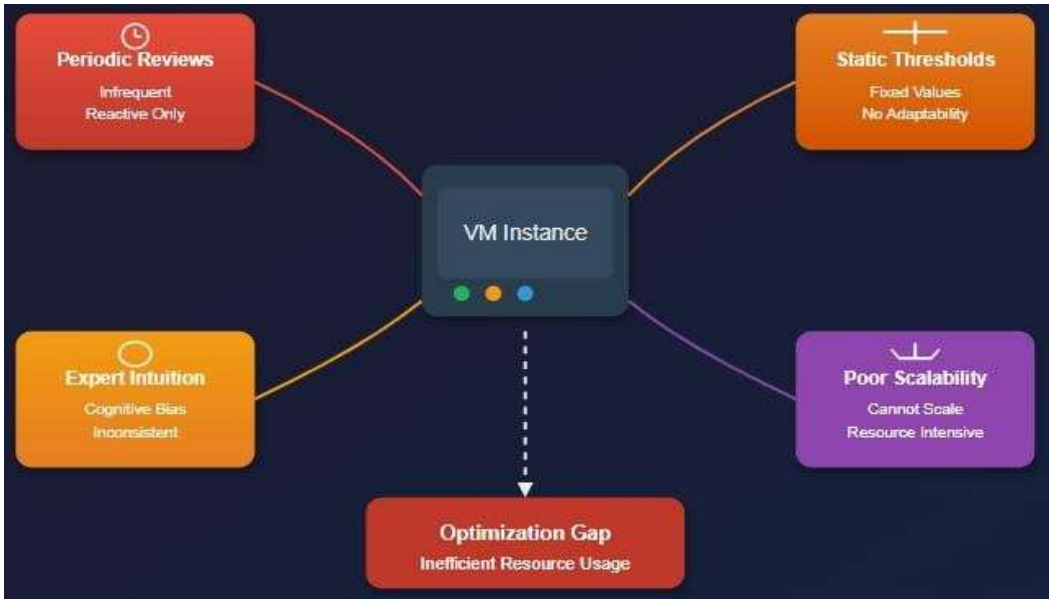

Fig 1: Manual VM Right-Sizing Limitations [3, 4]

Expert intuition, while valuable in navigating complex systems, introduces inconsistency and cognitive bias into right-sizing decisions. Engineers evaluating identical utilization data may arrive at different conclusions based on individual experience, risk tolerance, or mental models of system behavior. In

organizations operating across multiple data centers or cloud regions, this variability can result in inconsistent resource allocation for identical workloads, undermining governance consistency and complicating capacity planning at scale. Moreover, expert-driven analysis does not scale effectively as VM inventories grow. Human reviewers are unable to maintain holistic awareness across thousands of instances, particularly when optimization opportunities arise from aggregate patterns rather than individual VM behavior. The combination of infrequent reviews, static thresholds, and subjective judgment creates an optimization gap that manual methods cannot close, regardless of the level of human effort applied.These limitations motivate the exploration of automation, but as discussed in the next section, fully autonomous right-sizing introduces its own set of operational risks.

3. Operational Risks of Fully Automated Right-Sizing Systems

The limitations of manual right-sizing approaches have motivated the development of fully automated systems capable of analyzing utilization data and executing configuration changes without human intervention. These systems offer clear advantages in speed and scalability. However, when deployed in production environments, they introduce significant operational risks that require careful consideration. A study presented at the USENIX Symposium on Operating Systems Design and Implementation shows that many catastrophic failures in distributed systems stem from improper error handling, particularly when automated components encounter unexpected conditions and respond in ways that amplify rather than mitigate failures [5]. This finding has direct implications for automated infrastructure management, where right-sizing actions may interact unpredictably with other system processes, especially during abnormal workload behavior or when multiple automated mechanisms operate concurrently.

One of the most severe failure modes of fully automated right-sizing systems arises from aggressive resizing decisions. Cost-optimization algorithms may recommend substantial resource reductions based on periods of low observed utilization, without adequately accounting for workloads that exhibit periodic or burst-driven demand. Analysis of Google cluster traces published by the ACM Symposium on Cloud Computing demonstrates that cloud workloads are highly heterogeneous and dynamic, with resource requirements varying significantly across time horizons and workload types [6]. An automated system that evaluates utilization over a limited observation window may downsize resources during low-demand periods, only to encounter insufficient capacity when workload behavior shifts. The same analysis shows that short-term workload behavior is often not reliably predictable from historical data alone, as applications can change their resource consumption patterns abruptly. This limits the ability of fully automated systems to anticipate future demands using retrospective telemetry.

Customer expectations and service level agreements introduce an additional layer of complexity that automated systems are generally unable to capture. Many enterprise workloads are governed by contractual performance requirements related to latency, availability, or peak capacity, which are not directly inferable from utilization metrics. In such cases, resource allocations may intentionally exceed observed usage to satisfy business or compliance obligations. Automated systems that lack access to this contextual information risk generating inappropriate recommendations, which, if executed automatically, can result in SLA violations or customer dissatisfaction. These shortcomings are most evident in edge cases and exceptional scenarios, precisely where infrastructure decisions carry the highest risk and the cost of failure is greatest. Production environments therefore require operational approaches that can accommodate uncertainty, incorporate external context, and adapt to unexpected conditions capabilities that fully autonomous right-sizing systems, operating without human oversight, do not yet possess.

Risk Category	Description	Potential Consequence
Error Handling	Incorrect response to unexpected conditions	Cascading system failures
Aggressive Resizing	Over-optimization for cost efficiency	Performance degradation
Workload Heterogeneity	Inability to handle diverse patterns	Inappropriate resource allocation
SLA Compliance	No awareness of contractual obligations	Service agreement violations
Edge Cases	Poor handling of exceptional scenarios	Critical system instability
Contextual Blindness	Missing business and customer context	Customer dissatisfaction

Table 1: Operational Risks of Fully Automated Right-Sizing Systems [5, 6]

4. AI-Driven Workload Analysis and Recommendation Systems

AI-driven workload analysis systems serve as a pragmatic bridge between manual and fully automated right-sizing approaches. They leverage machine learning to interpret ambiguous and high-volume telemetry data while deliberately preserving human control over consequential decisions. These systems continuously ingest utilization metrics, performance signals, and operational telemetry from VM instances to construct behavioral models that capture temporal trends, correlations, and resource dependencies. This capability is essential in cloud environments where performance behavior is influenced by multiple interacting factors, including hardware heterogeneity, resource contention, and virtualization overhead [7]. Such complexity makes static rules and threshold-based approaches insufficient, motivating the use of AI to surface patterns and optimization opportunities that are not readily observable through conventional monitoring tools.

Within this framework, AI systems generate right-sizing recommendations as structured decision artifacts rather than automated actions. When a potential optimization opportunity is detected, the system produces a proposal that includes the suggested configuration change, quantified confidence indicators, estimated cost impact, and an assessment of potential performance risk, supported by historical evidence. This design explicitly separates analysis from execution, enabling scalability in evaluation without exposing production environments to ungoverned automation. Insights from intelligible model design in high-stakes domains such as healthcare demonstrate that AI systems can achieve both analytical accuracy and interpretability when explainability is treated as a core design requirement rather than an afterthought [8]. Infrastructure management shares similar characteristics, including high failure costs, the need for human supervision, and strong accountability requirements, making these design principles directly applicable.

Journal of Information Systems Engineering and Management
2026, 11(1s) e-ISSN: 2468-4376 https://www.jisem-journal.com/ **Research Article**

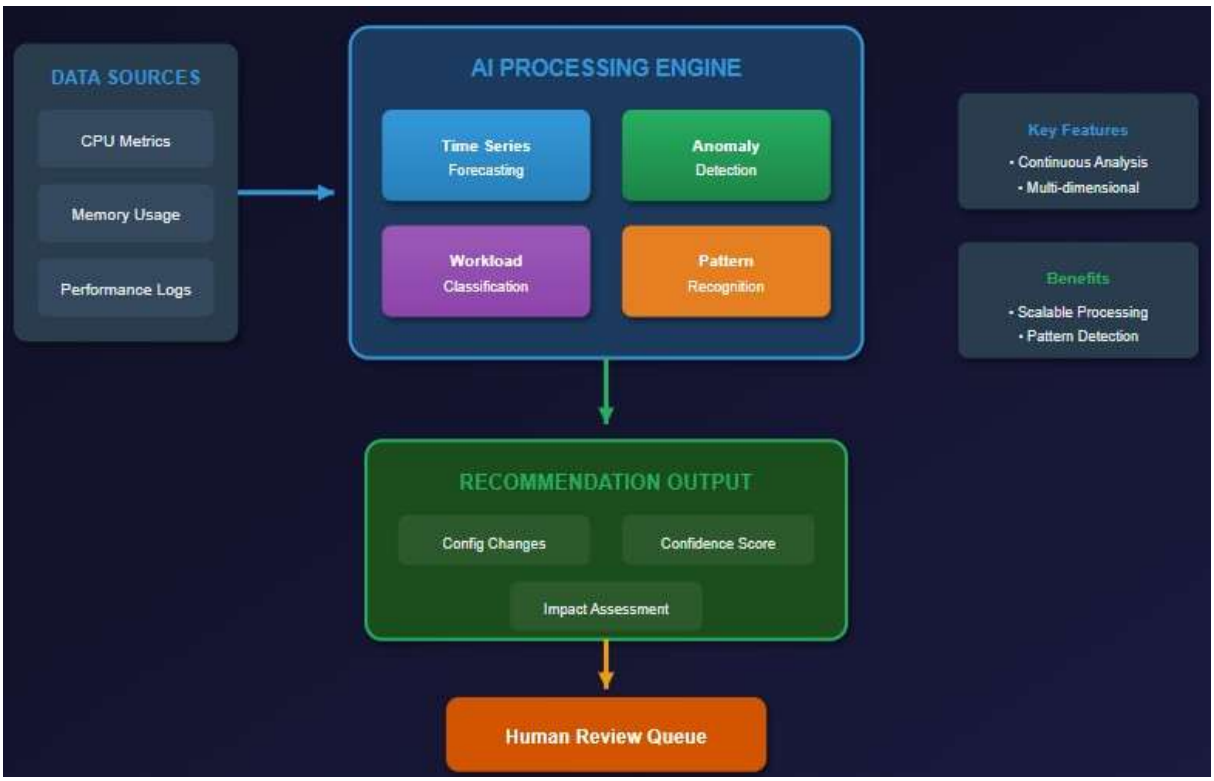

Fig 2: AI Recommendation System Architecture [7, 8]

From an architectural perspective, these systems rely on multiple specialized models that address distinct aspects of workload behavior, coordinated within a unified decision-support pipeline. Timeseries forecasting models are used to anticipate future resource demand and reduce the risk of premature downsizing ahead of workload growth. Anomaly detection components identify deviations from expected behavior that may indicate application issues, configuration drift, or legitimate workload transitions requiring investigation before optimization. Classification models are employed to distinguish between workload archetypes with fundamentally different resource characteristics, such as batch-oriented processing, interactive services, and stateless applications, enabling resizing strategies that are aligned with behavioral intent rather than raw utilization alone. By combining these capabilities into an ensemble architecture, the system can perform fine-grained analysis across heterogeneous workload portfolios while maintaining a conservative operational posture. Crucially, the output of this analysis remains advisory, allowing AI-driven insights to scale across large VM inventories without removing human authority over production changes.

5. Human-in-the-Loop Governance and Approval Frameworks

Human-in-the-loop governance establishes the organizational and technical structures through which AI-generated recommendations are reviewed, approved, and managed by humans before implementation. This governance model recognizes that, regardless of analytical sophistication, AI systems lack the contextual awareness, stakeholder relationships, and accountability that human decision-makers bring to infrastructure management. Research on theory-driven, user-centric explainable AI emphasizes that effective human–AI interaction depends on systems providing explanations that are relevant to the user's needs and the decision context [9]. In the domain of infrastructure management, this translates into equipping operators with sufficient insight into the rationale behind recommendations, enabling them to assess proposed actions and make informed

judgments about whether to proceed. Explanation requirements vary based on the reviewer's level of expertise, the sensitivity of the affected workload, and the magnitude of the proposed change.

Under human-in-the-loop models, platform operators acting as approval authorities apply multiple evaluation criteria when reviewing AI recommendations. Technical evaluation assesses whether a proposed change aligns with system architecture, capacity planning objectives, and established engineering standards. Contextual evaluation incorporates information beyond the analytical scope of the AI system, such as upcoming business events that may affect demand, application-specific sensitivities known to stakeholders, or customer communications signaling shifts in usage patterns. Risk evaluation weighs the potential benefits of optimization against the possible consequences of action, recognizing that not all technically sound recommendations are appropriate to implement. Research published by the ACM CHI Conference on explainable, accountable, and intelligible systems highlights the importance of calibrated trust in AI systems, enabling users to rely on system guidance when justified while remaining appropriately skeptical in uncertain situations [10]. This calibration is critical to effective human-in-the-loop governance, as reviewers must avoid both indiscriminate acceptance and reflexive rejection of AI-generated advice.

Fig 3: Human-in-the-Loop Governance Framework [9, 10]

Customer-specific considerations represent one of the most important dimensions of human oversight that automated systems cannot adequately capture. Enterprise cloud platforms serve a diverse customer base with varying risk tolerances, performance expectations, and contractual obligations. A recommendation that is suitable for an internal development workload may be inappropriate for a customer-facing production system governed by strict service level agreements. Human reviewers contribute relationship knowledge, contractual awareness, and organizational priorities that ensure optimization decisions align with customer expectations and business commitments. Effective approval frameworks must therefore support escalation paths for recommendations affecting sensitive workloads, enabling oversight proportional to the level of risk. This governance structure establishes clear accountability between AI recommendations and human decision-makers, supports responsible handling of unforeseen outcomes, and enables continuous improvement of AI models through reviewer feedback and observed operational results.

6. Future Operating Models for AI-Assisted Infrastructure Management

The evolution of AI-enabled infrastructure management is moving toward operating models that dynamically balance automation with human supervision, based on workload characteristics, organizational maturity, and risk tolerance. Future systems are likely to adopt tiered automation, in which routine optimization for well-understood workloads proceeds with minimal human involvement, while recommendations affecting new or sensitive systems require explicit approval. This tiered approach recognizes that human oversight, while essential, is a limited resource and should be focused on decisions where human judgment adds the greatest value. As automation increases, transparency requirements will also intensify. Operators must understand system behavior well enough to trust automated functions and to detect potential failure modes before they affect production environments. Research on energy-conscious resource allocation further suggests that future systems will need to optimize across increasingly complex and competing objectives, including performance, cost, and energy efficiency [1].

To support these operating models at scale, new forms of federated governance are likely to emerge. Such frameworks accommodate the organizational complexity of modern cloud deployments by enabling coordination across teams, business units, and geographic regions while preserving local autonomy and accountability. Platform teams can define global policies that specify acceptable optimization boundaries, risk thresholds, and escalation paths, while application teams retain the ability to make workload-specific decisions within those constraints. The Microsoft Research guidelines on human–AI interaction provide relevant design principles for these collaborative systems, emphasizing the importance of clearly communicating system capabilities and limitations and supporting effective correction when errors occur [2]. Federated governance extends the reach of human oversight by distributing review responsibilities across the organization while maintaining consistency through shared policy structures and centrally managed AI platforms.

Continuous-learning architectures further enhance the effectiveness of AI-assisted infrastructure management over time. By incorporating feedback from human reviewers and observing the outcomes of implemented changes, AI systems can progressively refine their models and recommendations. When a reviewer rejects a recommendation, the rationale can be captured and used to adjust future behavior, aligning system output more closely with organizational preferences and operational realities. Similarly, monitoring the impact of approved changes enables validation of predicted outcomes and ongoing refinement of forecasting models. These feedback loops create positive reinforcement cycles in which AI capabilities improve with operational experience, guided by human judgment and oversight. The long-term vision is a human–AI partnership in which technology provides scalable analysis and optimization, while humans contribute judgment, accountability, and contextual awareness, together enabling reliable, efficient, and sustainable infrastructure operations.

Conclusion

Human-supervised, AI-assisted VM right-sizing represents a mature and practical response to the limitations of both manual optimization and fully automated infrastructure management. In modern cloud environments, manual approaches cannot scale to meet operational demands, while fully autonomous systems introduce unacceptable risk in production settings where failures carry high impact. AI-assisted right-sizing, combined with human-in-the-loop governance, offers a balanced operating model that leverages the speed and scale of machine learning while preserving human judgment, accountability, and contextual awareness in decision-making.The technical foundations of this approach are well established. Contemporary recommendation systems are capable of processing large volumes of telemetry data and producing actionable guidance supported by confidence measures

and impact assessments. However, realizing the full benefits of this model requires corresponding organizational investment. Effective adoption depends on clearly defined approval frameworks, welltrained reviewers, and alignment with existing operational processes and accountability structures.When these elements are in place, organizations can achieve sustained improvements in cost efficiency and resource utilization without compromising reliability or customer trust. More broadly, the collaborative paradigm presented in this work positioning AI as an augmentative partner rather than an autonomous decision-maker provides a blueprint for responsible AI adoption across infrastructure operations beyond VM right-sizing, particularly in domains where scale, risk, and accountability must be carefully balanced.

References

[1] Anton Beloglazov et al., "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," ScienceDirect, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167739X11000689

[2] Saleema Amershi et al., "Guidelines for Human-AI Interaction," Microsoft, 2019. [Online]. Available: https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/

[3] Rahul Ghosh and Vijay K. Naik, "Biting off Safely More than You Can Chew: Predictive Analytics for Resource Over-commit in IaaS Cloud". [Online]. Available: https://www.researchgate.net/profile/Rahul-Ghosh-5/publication/261526165

[4] Qi Zhang et al., "Cloud computing: state-of-the-art and research challenges," ResearchGate, 2010. [Online]. Available: https://www.researchgate.net/publication/225252747_Cloud_computing_stateof-the-art_and_research_challenges

[5] Ding Yuan et al., "Simple Testing Can Prevent Most Critical Failures: An Analysis of Production Failures in Distributed Data-Intensive Systems," USENIX, 2014. [Online]. Available: https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-yuan.pdf

[6] Charles Reiss et al., "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," ACM, 2012. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/2391229.2391236

[7] Philipp Leitner and Juergen Cito, "Patterns in the Chaos - a Study of Performance Variation and Predictability in Public IaaS Clouds," arXiv:1411.2429, 2016. [Online]. Available: https://arxiv.org/abs/1411.2429

[8] Rich Caruana et al., "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," ACM, 2015. [Online]. Available: https://www.microsoft.com/enus/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf

[9] Danding Wang et al., "Designing Theory-Driven User-Centric Explainable AI," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/330967106_Designing_TheoryDriven_User-Centric_Explainable_AI

[10] Ashraf Abdul et al., "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda," ACM, 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3173574.3174156