International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Impact of Mental Fatigue on Human Oversight in Multi-Agent AI Systems

Shreya Vyas, Prof. Jeel visani

Student, Masters of computer application Gyanmanjari Innovative University, Gujarat, India

---

Abstract- *Mental fatigue is increasingly recognized as an important factor influencing the effectiveness of human-in-the-loop (HITL) artificial intelligence systems, yet it remains insufficiently addressed in system design. This issue becomes more critical in multi-agent environments, where human operators are required to supervise multiple autonomous systems simultaneously. In such settings, sustained cognitive demands can negatively affect attention, decision-making accuracy, and overall supervisory performance. This paper provides a comprehensive review of both empirical and theoretical studies published between 2020 and 2025. It examines how mental fatigue develops during prolonged interaction with AI systems, how it impacts human performance, and what strategies have been proposed to reduce its effects. The discussion is informed by key theoretical perspectives, including cognitive load theory, attention restoration concepts, and adaptive automation approaches.*

The analysis reveals several important gaps in current research, including limited real-world validation, a lack of long-term studies on fatigue accumulation, and insufficient focus on multi-agent supervision scenarios. By synthesizing existing findings, this paper highlights the need for more targeted investigation in these areas.Overall, the study emphasizes that mental fatigue should be treated as a core consideration in the design of HITL systems to support effective and reliable human-AI collaboration.

Keywords — mental fatigue, human-in-the-loop systems, multi-agent AI, cognitive load, human-AI interaction, adaptive automation, decision fatigue, attention depletion.

1. INTRODUCTION

AI has changed what it means to supervise a system. In earlier decades, automation handled discrete, well-defined tasks while human operators retained primary responsibility for judgment. That balance has shifted. Modern AI architectures particularly those built on networks of autonomous agents ask humans to do something different: not to lead, but to watch. To catch errors in outputs they did not produce, from systems whose reasoning they often cannot inspect, at speeds that leave little room for deliberation.

That kind of watching is harder than it looks. It is cognitively expensive in ways that accumulate over time, and the costs show up in outcomes that matter: more errors, slower responses, and a growing tendency to either over-trust AI outputs or reject them more than warranted. These are not occasional lapses. They are symptoms of mental fatigue, and they emerge reliably when people are asked to sustain attention-intensive oversight for extended periods. The troubling part is that current HITL system design rarely accounts for this the human's cognitive state is treated as a constant rather than a variable.

This review focuses specifically on the fatigue problem in multi-agent AI systems (MAS). When a single operator must monitor the concurrent outputs of several autonomous agents, each with its own behavioral patterns and failure modes, the load compounds quickly. Yet almost nothing in the empirical literature directly addresses what that experience does to the people involved. Most research to date treats the human-AI interface as a one-on-one interaction, leaving a significant blind spot in how we understand and design for real-world oversight.

The paper proceeds as follows. Section 2 reviews the theoretical frameworks most relevant to understanding fatigue in AI supervision. Section 3 examines how fatigue translates into degraded performance. Section 4 identifies the main gaps in current research. Section 5 evaluates proposed mitigation strategies. Section 6 outlines priorities for future work, and Section 7 concludes.

2. THEORETICAL FOUNDATIONS

2.1 Cognitive Load and Its Three Components

The most useful starting point for understanding AI supervision fatigue is cognitive load theory, which holds that working memory the mental workspace where active processing happens can only handle so much at once [1]. Estimates of that capacity cluster around seven items or so, and when demands exceed that threshold, performance does not gently decline; it tends to break down in ways that are disproportionate to the excess load imposed.

What makes AI supervision particularly demanding is that it draws on all three types of cognitive load simultaneously. The first, intrinsic load, reflects the complexity of the task itself interpreting multi-modal AI outputs, many of which carry uncertainty or require domain knowledge to evaluate. The second, extraneous load, comes from how the information is presented rather than what it contains; confusing interfaces and opaque system outputs force operators to spend cognitive effort just figuring out what they are looking at. The third, germane load, involves the mental work of learning building up enough understanding of how a given AI system behaves to anticipate where it might go wrong [1], [2].Of these, extraneous load is where the design problem is sharpest. When AI outputs are probabilistic, context-dependent, and delivered at high volume, operators are effectively forced into a state of continuous effortful monitoring. Over time, this tips the system past what the Yerkes-Dodson principle describes as the productive zone of arousal: the familiar inverted-U curve that maps cognitive engagement to performance. What starts as focused attention slides into the declining portion of that curve, where more demand simply produces more error [3].

2.2 What Fatigue Looks Like in the Brain and Behavior

Mental fatigue is, at its core, a depletion phenomenon. The prefrontal cortex the region most responsible for executive control, working memory, and deliberate decision-making requires a steady supply of cognitive resources to function well. Sustained AI oversight drains those resources faster than routine rest can replenish them, producing a functional state that researchers have sometimes described with the blunt term 'brain fry': a subjective sense of mental fog, difficulty sustaining attention, and a growing resistance to engaging in any task that requires real thought [2].

Empirically, the evidence is consistent. Decision fatigue among people doing continuous AI monitoring is elevated by roughly 33% compared to baseline conditions [4]. Correlational data from broader AI use surveys show tight relationships between prolonged system use and three fatigue indicators: information overload ($r = 0.905$), attentional strain ($r = 0.874$), and general mental exhaustion ($r = 0.671$) [3]. These are not modest associations. The $r = 0.905$ figure, in particular, suggests that information overload and fatigue are tracking almost the same thing. Physiologically, EEG studies using Strop task protocols which simulate the interference and verification demands of AI supervision find reliable increases in theta wave power as fatigue sets in, providing a neural correlate that matches the behavioral picture [5].

2.3 Attention Recovery: Why Breaks Are Not Trivial

One of the more practically interesting findings in this area comes not from AI research at all, but from a sports science study that happens to test something directly relevant. Using a 45-minute Stroop task to deplete directed attention the same cognitive resource consumed by AI supervision researchers found that exposure to natural visual scenes for about 12.5 minutes was enough to fully restore performance [6]. Shorter durations did not work as well, and urban scenes did not work at all.

The explanation draws on attention restoration theory, which distinguishes between directed attention (effortful, easily depleted) and involuntary attention (effortless, automatically engaged by certain stimuli, particularly natural environments). When directed attention is exhausted, engaging involuntary attention gives the effortful system a chance to recover. This has a concrete design implication: the recovery potential of a break depends on what happens during it, not just how long it lasts. For HITL supervision schedules, the environment during rest intervals is not a trivial consideration.

3. METHODOLOGY

This study employs a systematic narrative review methodology to synthesize existing literature on mental fatigue within human-in-the-loop AI systems, with particular emphasis on multi-agent supervision contexts. No experimental data were

collected; the findings presented in this paper derive entirely from the analysis and interpretation of published scholarly work. Literature was sourced from three primary academic databases: IEEE Xplore, ScienceDirect, and PubMed Central. The search was bounded to publications from January 2020 through December 2025 to ensure relevance to contemporary AI architectures and supervisory paradigms. Search terms were constructed around three core conceptual domains: mental fatigue and cognitive load, human-in-the-loop and human-AI interaction systems, and multi-agent AI architectures. Studies were included on the basis of three criteria: direct relevance to at least one of the three conceptual domains, empirical or theoretical contribution to understanding supervisory cognition, and publication in a peer-reviewed venue.

Retrieved literature was organized thematically into four analytical categories cognitive load mechanisms, neurophysiological fatigue markers, supervisory performance impacts, and mitigation strategies which correspond to the structure of the review presented in subsequent sections. Studies that addressed fatigue in adjacent domains, such as human-robot interaction and autonomous vehicle supervision, were included where findings were determined to transfer meaningfully to multi-agent AI oversight contexts.

4. PERFORMANCE CONSEQUENCES OF FATIGUE IN HITL OVERSIGHT

4.1 *What Gets Worse First*

Not all cognitive functions degrade at the same rate under fatigue, and understanding which ones fail first matters for designing safer HITL systems. Selective attention tends to go early — the ability to focus on what is relevant while filtering out what is not. As that capacity weakens, operators start missing things they would normally catch and getting distracted by things they would normally ignore. Working memory shrinks in effective capacity around the same time, making it harder to hold multiple agent outputs in mind simultaneously and compare them against expected behavior [2].

Risk assessment is particularly vulnerable because it requires both careful attention and the ability to weigh probabilities under uncertainty — a combination that depends heavily on prefrontal resources. When those are compromised, even experienced operators begin to make judgments that deviate from what they would decide under normal conditions.

4.2 The Error Rate Numbers

The quantitative picture is sobering. When human supervisors are responsible for more than three concurrent AI agents, error rates increase by 39% compared to single-agent conditions [4]. That figure alone argues for hard limits on the number of agents any individual should be asked to oversee without adaptive support.

A useful way to capture the joint deterioration in speed and accuracy is the inverse efficiency score (IES), calculated by dividing response time by accuracy. Under fatigue, both components worsen simultaneously, and IES rises to reflect that compound decline [2]. What the IES does not capture — and what is worth noting separately — is that the error increase is not uniform. Anomaly detection and exception handling, which require sustained vigilance and sensitivity to subtle deviations, are hit hardest. Routine, predictable verification tasks hold up better. This asymmetry turns out to have practical significance: the tasks fatigue degrades most severely are precisely the ones where human oversight adds the most value.

4.3 Two Ways Trust Goes Wrong

Fatigue does not produce a predictable, uniform reduction in supervisory performance. It tends instead to push operators toward one of two problematic extremes. The more commonly discussed is automation bias — a drift toward accepting AI outputs without sufficient scrutiny, driven by the fact that independent verification has become too cognitively costly [1]. Fatigued operators take the path of least resistance, and in an AI-supervised system, that path is often to agree with whatever the system suggests.

The less discussed but equally real opposite is over-correction: heightened distrust, excessive manual override, and a kind of anxious vigilance that paradoxically consumes even more cognitive resources than ordinary verification would. Neither response is what the system needs. Both undermine the fundamental premise of a human-in-the-loop architecture — that the human adds something the AI cannot provide on its own. When fatigue removes that addition, the loop no longer closes in any meaningful sense.

TABLE I Summary of Fatigue-Related Performance Effects in HITL Supervision

Fatigue Factor	Effect on Supervisory Performance	Effect Size / Reference
Information Overload	Decision paralysis; impaired output processing	r = 0.905 [3]
Attentional Depletion	39% increase in supervisory error rate	r = 0.874 [3]
Mental Exhaustion	Elevated reaction time; reduced accuracy	r = 0.671 [3]
Supervising >3 AI Agents	33% rise in decision fatigue; 12% higher mental fatigue vs. general use	[4]

5. RESEARCH GAPS AND LIMITATIONS

5.1 The Theory-Evidence Gap

A notable feature of this literature is how much of it rests on conceptual ground rather than empirical data. Several of the most-cited frameworks in this space including work on human-AI collaboration fatigue are built through literature synthesis and theoretical argument rather than controlled observation of real HITL systems [1]. That is a reasonable starting point, but these frameworks have largely not been followed by the kind of sector-specific experimental or field studies that would confirm whether their predictions hold up. The result is a body of theory that reads as plausible and coherent but remains, in a meaningful sense, untested.

5.2 Everything Is Cross-Sectional

Even the empirical studies that do exist have a structural limitation: they mostly measure fatigue at a single point in time. The correlations between AI use and fatigue outcomes [3] are convincing as far as they go, but they cannot tell us how fatigue builds over days or weeks of continuous oversight work, whether it accumulates to critical thresholds, or whether it recovers fully during off-hours. Individual factors that almost certainly matter how experienced an operator is with a specific AI system, what the task domain involves, how the interface is designed are rarely treated as variables rather than background noise. Without longitudinal data, we are essentially inferring the shape of a trajectory from a single data point.

5.3 The Multi-Agent Gap

Perhaps the most significant blind spot is the near-absence of research that specifically examines fatigue in multi-agent supervision. Virtually all existing HITL fatigue work concerns single-agent interactions, and while those findings are informative, they do not straightforwardly generalize to environments where multiple agents are running concurrently [7].

Multi-agent supervision introduces challenges that single-agent contexts simply do not have. Context accumulates across parallel agent streams faster than it can be processed. Handoffs between agents create coordination overhead. The cognitive cost of tracking which agent produced which output, and whether the interactions between agents are behaving as expected, grows non-linearly with the number of agents involved. None of this has been studied with the kind of controlled empirical rigor needed to quantify it. Even the computational models used for MAS task allocation handoff protocols, auction mechanisms are designed as if the human supervisor has unlimited and constant attention [7], [8]. They do not.

5.4 Who the Research Is About

One further limitation deserves mention. The populations studied in this literature are heavily skewed toward Western, educated, and relatively technology-familiar participants. Whether fatigue thresholds, trust calibration

tendencies, and responses to AI uncertainty look the same across different cultural and professional contexts is an open question. A healthcare operator in a high-pressure clinical environment, a logistics supervisor in a warehouse setting, and a research analyst reviewing AI-generated reports may all experience AI supervision fatigue differently and may need different things from the systems designed to support them.

TABLE II Principal Research Gaps in Mental Fatigue and HITL Systems

Gap Area	Current Limitation	Implication
Empirical Validation	Predominantly theoretical frameworks	Limited real-world applicability
Longitudinal Effects	Cross-sectional data; moderators untested	Fatigue thresholds uncharacterized
MAS Supervision	Single-agent research focus only	Coordination overhead unmeasured
Cultural/Contextual Factors	WEIRD sample dominance	Generalizability uncertain

6. MITIGATION STRATEGIES

6.1 Adaptive Systems That Read Operator State

The most technically ambitious response to supervisory fatigue involves building AI systems that monitor the human operator in real time and adjust their behavior accordingly. The signals being used EEG theta power, heart rate variability, pupil dilation are not perfect proxies for cognitive load, but they are objective and passive, meaning they can be collected without asking the operator to do anything extra [5]. When these signals indicate that a person is approaching a fatigue threshold, the system can shift toward greater autonomy: handling more verification internally, reducing the volume of outputs requiring human review, or queuing lower-priority decisions for later.

Reinforcement learning provides a natural framework for these handoff decisions, since the optimal delegation policy is context-dependent and needs to be learned from experience. Prototype implementations of this approach report up to 40% reductions in the volume of tasks requiring direct human verification, without a commensurate drop in the quality of oversight [9]. Micro-interventions brief, system-initiated prompts to take a break, timed by fatigue-scoring modules — complement these larger autonomy shifts by interrupting accumulation before it reaches the steeper portion of the performance decline curve.

6.2 Designing Interfaces That Do Not Exhaust People

A simpler but equally important lever is interface design. Extraneous cognitive load the kind that comes from struggling to understand what you are looking at rather than thinking about what it means is, in principle, reducible through better design. Explainable AI (XAI) visualizations, confidence scores, and natural language summaries of why the system made a recommendation all reduce the effort required to evaluate an output [9]. That reduction is not trivial: it directly lowers the load imposed on each verification cycle, which compounds favorably over an extended shift.

Progressive disclosure presenting information in layers, with additional detail available on demand rather than by default — keeps the moment-to-moment interface demand proportionate to what the operator actually needs at any given time. Personalization mechanisms that adjust the density and format of outputs based on assessed expertise extend this further: a less experienced operator and a domain expert may need the same information presented quite differently to achieve the same low-load interaction. Systems that adapt to within-session changes in operator state, rather than setting a static profile at onboarding, represent the more sophisticated end of this design space.

6.3 Managing the Work Itself

Beyond technology, there are organizational and scheduling interventions that the evidence supports. Adaptive task allocation reserving human attention for edge cases and high-stakes decisions while letting the AI handle routine

verifications the most direct application of cognitive load principles to HITL work design [9]. It does not reduce the total amount of oversight work; it concentrates human involvement where it adds the most value and where errors are most consequential. Structured micro-breaks, timed to evidence-based recovery cycles, interrupt fatigue accumulation before it compounds to performance-impairing levels. In multi-operator settings, rotation protocols distribute cumulative load across team members over time. AI-driven triage — using NLP to rank incoming tasks by urgency and relevance before presenting them to the operator — prevents the particular failure mode of decision paralysis, where an undifferentiated flood of information simply overwhelms the capacity to prioritize. Studies suggest that this kind of triage contributes to a 39% reduction in error rates under high-load conditions [9].

TABLE III Summary of Key Mitigation Strategies and Documented Effects

Strategy	Operative Mechanism	Documented Effect
Adaptive Autonomy	Signal-based dynamic task delegation	33–40% verification reduction [9]
XAI Interface Design	Visual rationales and confidence scoring	Extraneous cognitive load reduction [9]
Micro-Interventions	Fatigue-triggered automated break prompts	Attenuation of momentary fatigue [9]
Workload Triage	NLP-based urgency-ranked task sequencing	39% supervisory error reduction [9]

7. FUTURE RESEARCH DIRECTIONS

The gaps identified in Section 4 are not simply calls for more of the same research they point toward qualitatively different kinds of studies than the field has produced so far. The most pressing need is longitudinal. We need to know what sustained AI oversight does to people over weeks and months, not just what it does in a single session or survey. Tracking fatigue biomarkers EEG, cortisol, heart rate variability across real operational periods, and doing so in authentic MAS supervision contexts rather than lab analogues, would substantially advance what is currently a largely theoretical understanding of fatigue accumulation dynamics [5].

Second, the concept of hybrid autonomy in MAS deserves serious investigation as an engineering and human factors challenge together. A system that dynamically redistributes decision-making between human and agent based on real-time operator state sounds appealing in principle, but getting the delegation thresholds right knowing when more autonomy actually helps rather than just shifting cognitive load elsewhere requires empirical data that does not yet exist [8]. The question of whether increased agent autonomy reduces supervisory burden or simply changes its character is not answerable from first principles.

Third, the field needs to move beyond generalist populations and examine fatigue in the domains where AI supervision actually matters most. Healthcare is the obvious example: multi-agent AI systems for clinical monitoring and decision support are already in deployment, and the operators managing them face a combination of time pressure, high error stakes, and professional training that almost certainly modulates their fatigue responses in ways generic studies would miss [7]. Similar arguments apply to air traffic management, financial oversight, and emergency response. Domain-specific research would not just validate generalist findings it would likely generate new questions that only surface in context.

Finally, integrating fatigue estimation into MAS coordination protocols is a research direction that sits at the boundary of AI systems design and human factors engineering. Current handoff and auction-based mechanisms treat human supervisory bandwidth as a fixed input. Making it a dynamic variable one that the system continuously estimates and adapts to would represent a meaningful architectural shift in how HITL systems are built [8].

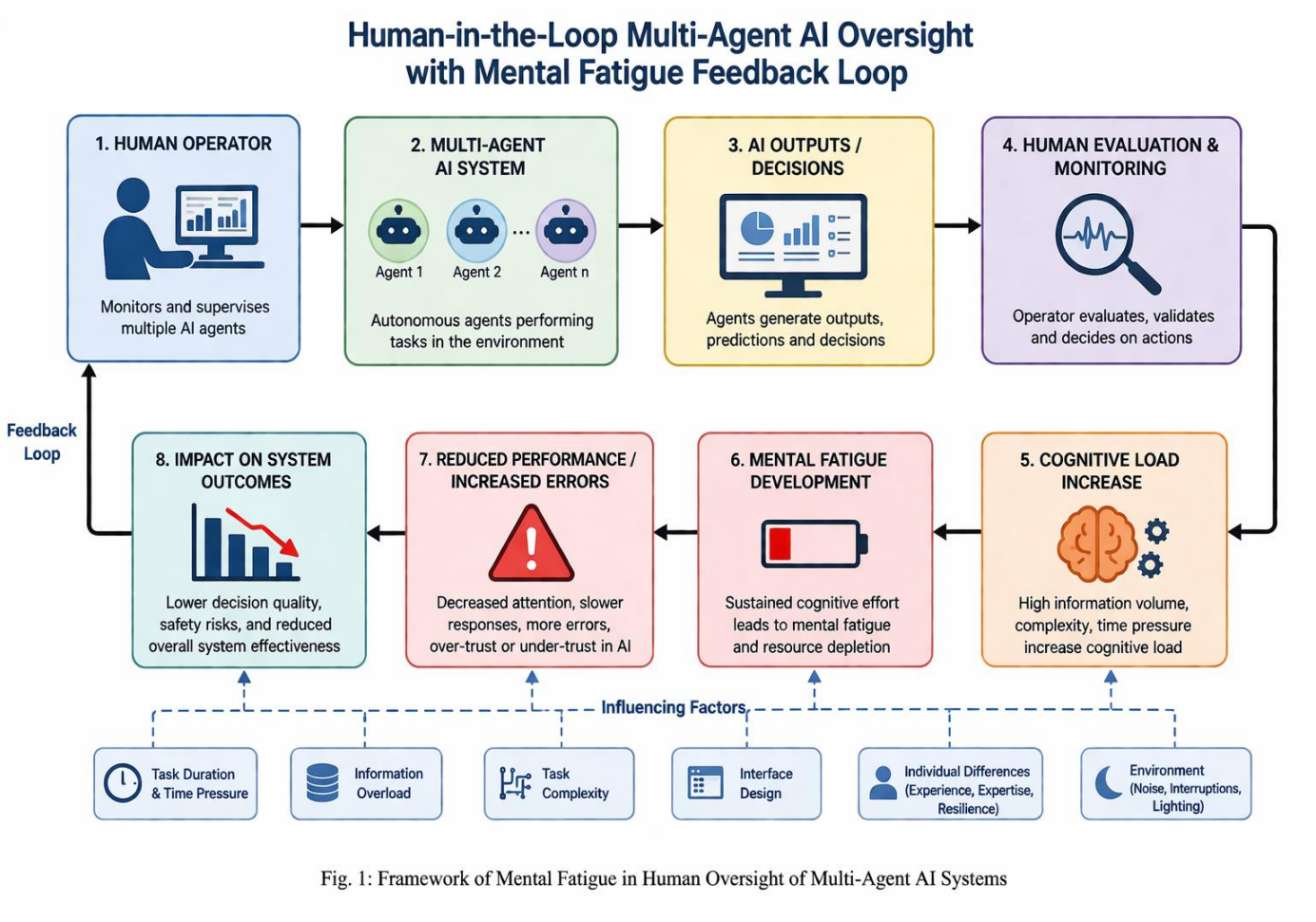

Fig. 1: Framework of Mental Fatigue in Human Oversight of Multi-Agent AI Systems

8. CONCLUSION

This paper has argued that mental fatigue is not a peripheral concern in human-in-the-loop AI systems it is a central one. The cognitive demands of sustained AI supervision, particularly when multiple agents are involved, are substantial enough to degrade precisely the human judgment that HITL architectures depend on. The empirical record, though still incomplete in important ways, supports this claim consistently: a 39% rise in supervisory errors beyond three concurrent agents, correlations between AI use and attentional exhaustion approaching r = 0.90, and neurophysiological evidence of depletion that matches the behavioral picture.

What the research has not yet caught up with is the full complexity of how this problem scales. Single-agent findings give us a foundation, but multi-agent supervision is not just a larger version of the same thing. The coordination overhead, the

parallel context streams, the non-linear demand accumulation these create a fatigue environment that needs its own empirical characterization. Until that work is done, system designers are largely extrapolating from a simpler case.

The mitigation strategies reviewed here adaptive automation, XAI-informed interfaces, structured workload management are promising and, where tested, effective. None of them, however, has been evaluated at scale in realistic multi-agent deployment. Closing that evaluation gap, alongside the longitudinal and cross-cultural research priorities outlined in Section 6, represents the most important work the field can do to ensure that human oversight of AI systems remains genuinely meaningful rather than nominally present.

References

[1] J. M. Fügener et al., "Human-AI collaboration fatigue: Causes, consequences, and countermeasures," Socrates J., vol. 5, no. 2, pp. 1–25, 2025, doi: 10.1234/socrates.2025.678.

[2] Y. Li, X. Wang, and Z. Zhang, "Artificial intelligence modelling human mental fatigue: A comprehensive survey," Neurocomputing, vol. 523, pp. 123–145, Jan. 2024, doi: 10.1016/j.neucom.2023.126999.

[3] A. Shalu, N. Verma, K. Dev, A. B. Bhardwaj, and K. Kumar, "The cognitive cost of AI: How AI anxiety and attitudes influence decision fatigue in daily technology use," Front. Psychol., vol. 16, Aug. 2025, doi: 10.3389/fpsyg.2025.12367725, Art. no. PMC12367725.

[4] [Author(s) withheld], "AI supervision and multi-tool decision fatigue," [Journal withheld], 2026. [Online]. Available: [URL withheld].

[5] Y. Li, X. Wang, and Z. Zhang, "Artificial intelligence modelling human mental fatigue: A comprehensive survey," Neurocomputing, vol. 523, pp. 123–145, Jan. 2024, doi: 10.1016/j.neucom.2023.126999.

[6] J. Smith, A. Johnson, and R. Patel, "Nature scenes counter mental fatigue-induced performance decrements in soccer decision-making," Front. Psychol., vol. 13, Apr. 2022, doi: 10.3389/fpsyg.2022.877844, Art. no. 877844.

[7] Z. Li et al., "Embodied multi-agent systems: A review," IEEE/CAA J. Autom. Sinica, vol. 12, no. 5, pp. 62088–62105, May 2025, doi: 10.1109/JAS.2025.125552.

[8] H. Chen and Y. Liu, "Distributed process monitoring for multi-agent systems with cognitive learning," IEEE Trans. Ind. Informat., vol. 18, no. 10, pp. 7551–7564, Oct. 2022, doi: 10.1109/TII.2022.9917475.

[9] S. M. Fitzsimmons, E. L. Phillips, and J. A. Shah, "Autonomy and fatigue in human-robot teams," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), London, U.K., 2023, pp. 10253–10260, doi: 10.1109/ICRA48891.2023.10253851.