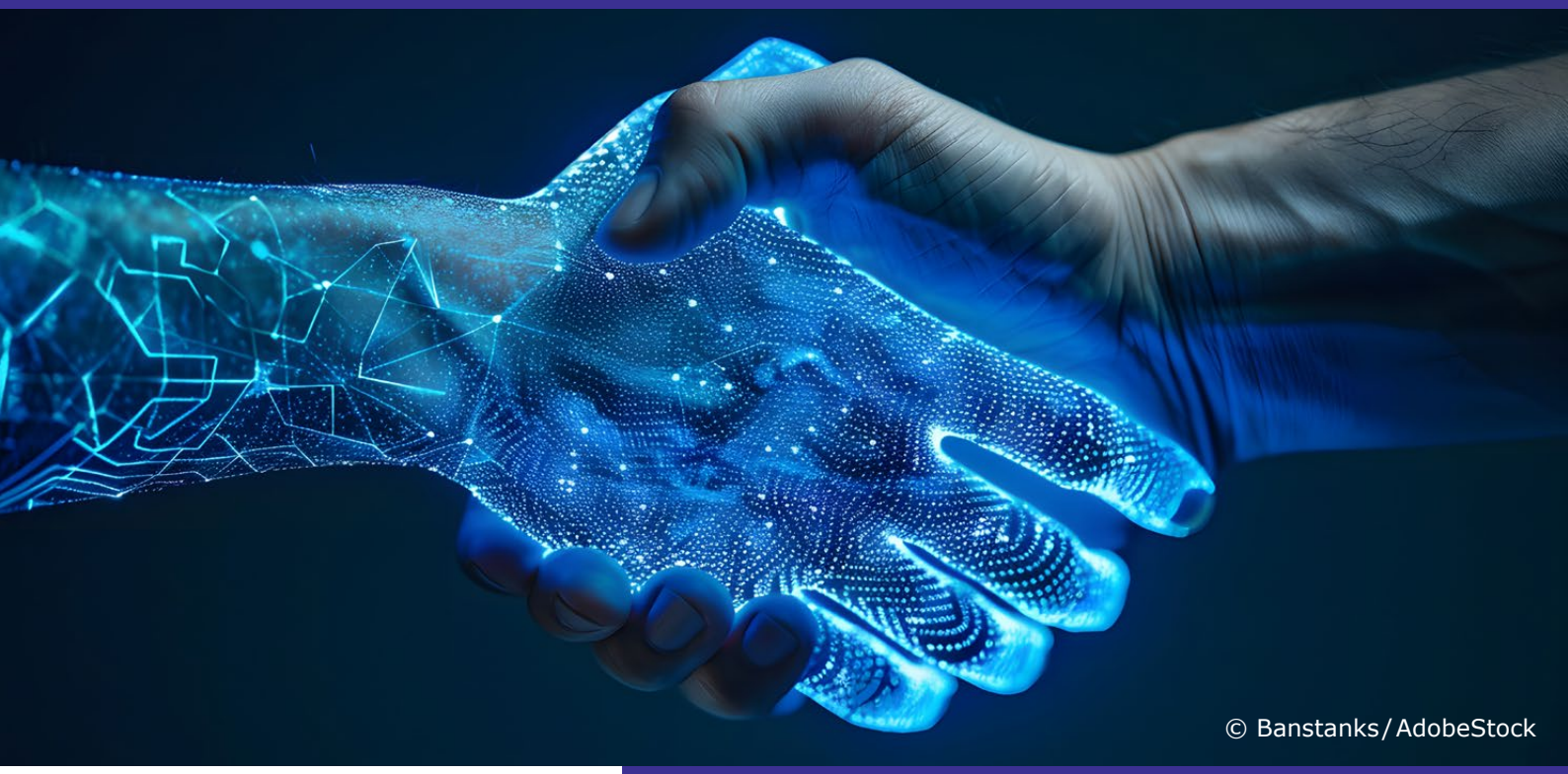
ESCP
BUSINESS SCHOOL

IT ALL STARTS HERE

BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW



LIGHTS - Technology

AI Proficiency: Redefining performance measurement for the human-AI hybrid era

ESCP Impact Paper No.2026-40-ENG

Erik STRAUSS
ESCP Business School

Christina STRAUSS
University of Duisburg-Essen

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ESCP Impact Paper No.2026-40-ENG

AI Proficiency: Redefining performance measurement for the human-AI hybrid era

Erik Strauss*
Christina Strauss**

ESCP Business School

Abstract

As generative artificial intelligence reshapes organizational life, legacy performance metrics rooted in individual output volume are producing a troubling "performance paradox", i.e., accelerating activity without necessarily advancing value. This paper argues for a fundamental reorientation of how organizations and educational institutions assess human contribution in AI-augmented environments. Drawing on the AI Proficiency framework and emerging team-level metrics such as the Human-Agent Ratio, it introduces practical tools for evaluating individual capability, team orchestration quality, and the effectiveness of human oversight. The argument extends to the educational sphere, where assessment must shift from measuring final outputs to measuring cognitive transformation. For managers, educators, and policy-makers, the paper proposes a three-pillar roadmap for the coming years: redefine metrics to capture synergy rather than speed, align incentive structures with genuine human-AI collaboration, and redesign education to cultivate the judgment, adaptability, and ethical awareness that machines cannot replicate.

Keywords : AI proficiency, performance measurement, human-AI collaboration, organizational learning, intelligent organization

*Professor, ESCP Business School
**University of Duisburg-Essen

AI Proficiency:
Redefining performance measurement for the human-AI hybrid era

The performance paradox of the Artificial Intelligence age

The transition of the global economy toward permanent human-AI hybridization constitutes the most significant rupture in performance measurement since the advent of scientific management. As organizations increasingly deploy generative artificial intelligence and autonomous agents across core functions, the frameworks traditionally used to evaluate human performance are becoming not merely outdated but actively counterproductive (Strauss, 2025). For most of modern economic history, performance measurement has been anchored in the logic of throughput and individual output volume, such as widgets produced, calls handled, and reports filed. In the age of AI-augmented work, this logic generates what we call the "AI performance paradox", i.e., organizations mistake the acceleration of tasks for the advancement of value.

The paradox is empirically visible. In marketing, AI systems can significantly increase lead volume while conversion rates simultaneously decline, because the system has been optimized for quantity over relational quality (e.g., Brynjolfsson et al., 2017). In software engineering, the long-standing metric of "lines of code" fails entirely to account for the verification overhead required to review and debug AI-generated outputs, which often introduce subtle errors that are significantly harder to detect than those a junior developer writing code manually would make (Ponnusamy, 2025). Evidence from recent field research further sharpens this pattern (Ranganathan & Ye, 2026): rather than reducing work, AI systematically intensifies it by expanding task scope, increasing parallel workflows, and raising implicit expectations for speed and output. Employees take on a broader range of responsibilities, revisit previously deferred tasks, and manage multiple AI-assisted processes simultaneously, which in turn generates additional layers of coordination, review, and cognitive load. This dynamic creates a self-reinforcing cycle in which higher output leads to higher expectations, and higher expectations drive further reliance on AI. As a result, what appears as productivity gains in traditional metrics often masks workload creep, cognitive fatigue, and declining decision quality, thereby distorting actual value creation. What is most consequentially eroded in this cycle is not speed or volume but what might be termed deliberative judgment: the slow, effortful, contextually sensitive reasoning that human beings perform when stakes are high, and information is ambiguous. This form of cognition, i.e., drawing on experience, ethical intuition, and situational awareness that resists algorithmic encoding, is precisely what AI workload creep displaces first. As cognitive bandwidth is consumed by the coordination and verification demands that AI intensification generates, the capacity for genuine deliberation shrinks. The performance paradox is therefore primarily a cognitive problem, raising the following question: if human beings are no longer valued for what they produce, what are they valued for? And how do we protect and develop that capacity?

What is needed is a shift from measuring individual output to measuring what we call "AI Proficiency", i.e., the quality of the synergy between human judgment and machine processing. This reorientation requires a (philosophical) rethinking of what human contribution means in an era where the execution of tasks can increasingly be delegated to intelligent systems. Human beings' distinctive contribution lies no longer in output volume but in the ability to direct, evaluate, correct, and contextually enrich the outputs that machines generate.

Redefining individual value: the AI Proficiency framework

To capture the true value of human-AI interaction at the individual level, we propose a multidimensional AI Proficiency framework organized around three core pillars (Strauss et al., 2026): (1) technical expertise, (2) domain knowledge and business awareness, and (3) organizational and human competencies. Together, these dimensions cover more than what superficial proxies (such as the number of prompts written or AI tools installed) assess the depth of integration between human capability and machine potential.

Technical expertise

True technical proficiency begins with a conceptual understanding of how (generative) AI functions, including the ability to recognize and explain phenomena such as hallucinations and to understand how variables like context length or temperature settings shape model outputs. Beyond this foundational literacy, highly proficient employees demonstrate operational skill by iteratively refining their inputs and tooling integration by connecting AI capabilities to existing organizational workflows through automation interfaces or APIs. The most advanced practitioners venture into prototyping, i.e., design lightweight evaluation tests to assess AI performance against specific business cases. This is a form of experimental thinking that transforms employees from passive users of AI tools into active architects of human-machine workflows.

Domain knowledge and business awareness

Technical capability without contextual judgment is insufficient, as even the most capable AI systems will fail to generate organizational value when deployed in misaligned or poorly understood business contexts. The second pillar of the framework, therefore, concerns employees' ability to translate AI capabilities into organizational value. This involves what the framework terms value identification, i.e., the capacity to clearly articulate a business problem in a way that makes it tractable for AI-assisted exploration, as well as feasibility awareness, the ability to assess whether a proposed use case is viable given the organization's data environment, risk appetite, and regulatory constraints. Leaders who demonstrate transformation awareness go further still, anticipating how AI will reshape roles over time and identifying the organizational and cultural barriers to adoption before they stall a project. This foresight transforms AI proficiency from an individual competency into a source of strategic advantage.

Organizational and human competencies

The third pillar addresses the mindset and competencies required to succeed in a hybrid work environment. Curiosity and learning orientation, i.e., the active exploration of new tools, the willingness to experiment and fail, and the practice of seeking feedback on one's AI applications, are foundational. Equally important is ethical reflection: the capacity to question whether AI outputs align with organizational values, whether they introduce hidden biases, and whether their use in a given context is appropriate. Finally, adaptability, i.e., the ability to remain confident and constructive when AI disrupts established routines, is itself a measurable skill. Employees who view these disruptions as opportunities for skill development rather than threats to their relevance, therefore, directly determine how effectively organizations can translate AI investments into sustained performance. What distinguishes these competencies from the technical and domain pillars is that they are structurally beyond AI. Ethical reflection, for instance, is not a matter of applying rules or

detecting policy violations, both of which AI can approximate with increasing sophistication. It involves contextual moral reasoning grounded in lived experience, institutional memory, and, crucially, personal accountability: the capacity to be held responsible for a decision and to bear its consequences. AI systems can flag ethical risk; they cannot own it. Similarly, adaptability in the human sense is not tolerance of parameter change but the capacity to reframe one's professional identity when the tools of one's vocation are disrupted – an existential act that requires agency. Another competency that deserves explicit recognition is relational intelligence, i.e., the ability to build trust, navigate conflict, read interpersonal dynamics, and sustain productive relationships under conditions of ambiguity and pressure. Research consistently shows that in high-stakes contexts, i.e., leadership transitions, cross-functional negotiations, client-facing decisions, the perception of genuine human presence, empathy, and accountability is itself a source of value that AI-mediated interaction cannot replicate (Dell'Acqua et al., 2025). Relational intelligence is a prerequisite for AI proficiency because it is the condition under which AI-generated outputs are accepted, trusted, and acted upon by other human beings.

Team-level metrics: from span of control to span of orchestration

While AI Proficiency measures individual capability, organizations equally need metrics to assess team-level orchestration and the safety of human oversight. As AI agents increasingly perform tasks that were previously distributed across multiple human roles, the classical managerial concept of "span of control" (i.e., the number of employees a manager supervises) gives way to what is better described as "span of orchestration": the number of AI agents and humans a human effectively coordinates and governs.

A critical emerging key performance indicator in this context is the "Human-Agent Ratio" (HAR), which represents the number of AI agents deployed per employee in a given team or function. In what researchers and practitioners are beginning to call "frontier firms," a higher HAR is associated with measurable capacity gains (Dell'Acqua et al., 2025; Microsoft, 2025). However, the HAR is not simply a number to maximize. In creative, strategic, or high-stakes decision-making contexts, a lower ratio (typically between 1:1 and 1:5) is often more effective. Beyond a certain threshold, organizations encounter what might be termed "agent chaos": agents generate outputs faster than human reviewers can evaluate them with appropriate rigor, producing the very output trap that AI was supposed to eliminate (Ranganathan & Ye, 2026).

Alongside the HAR, organizations should measure oversight effectiveness through an "Intervention Benefit Ratio" (IBR), i.e., a metric that quantifies how much human guidance improves AI accuracy in a given task domain. Research on human-AI collaboration in complex analytical tasks, such as diagnosis in medicine, demonstrates that a human-in-the-loop approach still outperforms standalone AI solutions (Olawade et al., 2026). The practical implication is clear: human oversight should be most intensively applied where its marginal contribution to decision quality is highest. Measuring this ratio allows organizations to deploy human attention strategically rather than reflexively, ensuring that the value of human judgment is not diluted by applying it indiscriminately across tasks where AI performs adequately on its own. To make this principle operational, organizations need a working typology of the task domains in which human contribution is not merely helpful but structurally irreplaceable. Three such domains deserve particular attention. The first is high-stakes moral decisions: situations in which an outcome carries significant consequences for people, communities, or institutions, and in which accountability cannot be legitimately delegated to a system that cannot be held responsible. Whether to restructure a team or communicate a difficult diagnosis, these are decisions that require a

human being capable of owning the outcome. The second domain is relational and trust-dependent interaction: contexts in which the perception of human presence is itself part of the value being delivered. In leadership, negotiation, client relationships, and therapeutic settings, people do not merely want an accurate output; they want to be heard, understood, and responded to by someone who can be genuinely affected by their situation. While AI can simulate this, it cannot perform it. And even if it could, the social and psychological acceptance of such substitution remains, for now, far from sufficient. The third domain is genuinely novel situations: circumstances for which no reliable training data exists, where pattern-matching fails, and where the appropriate response must be constructed from first principles, ethical intuition, and creative inference. These are precisely the conditions under which especially large language models are most unreliable and human judgment is most necessary. Calibrating the IBR with this typology in mind allows organizations to move from abstract aspiration to concrete deployment decisions about where human attention generates its greatest returns.

Incentivizing synergy: aligning rewards with human–AI partnership

New metrics are necessary but not sufficient. For the AI Proficiency framework, the HAR, and the IBR to drive meaningful behavioral change, they must be embedded in the reward structures that shape career trajectories. Organizations that have moved furthest in this direction use proficiency dimensions as formal rubrics in performance appraisals, with the explicit goal of rewarding what might be called “Human-AI Team Wins”, i.e., outcomes attributable to the quality of the human-machine collaboration rather than to either party alone.

Several early adopters illustrate this principle. Shopify has implemented peer-rating systems in which employees are assessed on how fluently and reflexively they use AI to amplify their individual contributions. Amazon’s robotics division has made demonstrating measurable AI-driven improvements in process outcomes a formal condition for promotion. These adaptations signal to the entire organization that mastering the capacity to work with AI (technically, strategically, and ethically) is the primary path to professional advancement. The cultural effect of this signal is arguably more powerful than the metrics themselves.

For organizations yet to make this shift, the risk is not merely competitive disadvantage but a form of organizational cognitive dissonance: asking employees to invest in developing AI collaboration skills while continuing to reward them on metrics that penalize the time needed for that investment. Aligning incentives with the capabilities that genuinely matter in the hybrid era is hence a strategic imperative.

Redesigning education for the Artificial Intelligence era

The metrics and incentive structures proposed above address a workforce that already exists. But the deepest challenge lies upstream: the educational system that produces and develops these employees remains largely designed for an industrial world that no longer exists. Just as organisations have defaulted to measuring output volume rather than the quality of human-AI synergy, educational institutions have inherited a parallel dysfunction, i.e., assessing students by what they produce, rather than by the quality of how they think. The same performance paradox that afflicts the modern workplace is thus reproduced in the lecture hall and the examination room before employees ever enter organisations.

Current educational models are, at their core, structured around the transmission of knowledge and the measurement of output per unit of time, i.e., precisely the logic that fails

in AI-augmented environments. Students are routinely assessed by putting them in direct competition with AI: generating essays, solving equations, and producing code under time pressure. This is a competition that human cognition is structurally designed to lose, and asking students to "beat" AI at output generation is to entirely misunderstand what education in the Artificial Intelligence Era should be for.

The core value of education must shift from knowledge transmission to cognitive stimulation, i.e., a reorientation that mirrors, at the individual developmental level, the shift from measuring throughput to measuring AI Proficiency at the organisational level. In an era where access to information is instantaneous, what education can uniquely provide is not content but the capacity to think with, against, and beyond that content. This means valuing active inquiry, generative problem-solving, and counterfactual reasoning over passive reception. It means cultivating what learning scientists call productive struggle: the metacognitive experience of persisting through ambiguity, revising one's approach, and developing genuine understanding rather than retrieved answers. And it means placing a higher premium on humanistic mastery: empathy, ethical judgment, curiosity, and the capacity to make decisions in the absence of clear algorithmic guidance. Take empathy as an example: AI systems can recognise emotional cues with increasing accuracy and generate responses that conform to empathic conventions. However, empathy in the morally and socially significant sense is the willingness to be affected by another person's reality, to allow their situation to alter one's own response rather than mere pattern recognition. That requires vulnerability (i.e., a capacity to be changed by an encounter), which is a feature of embodied, experiential consciousness. Ethical judgment involves a related structure: not the application of a decision rule to a case, but the capacity to sit with genuine moral uncertainty, to acknowledge competing obligations without resolving them prematurely, and to act anyway under conditions of irreducible ambiguity. This gap in current AI capability will not be closed by more training. It is a constitutive difference in what kind of entity can bear moral responsibility and accountability. Education that takes these distinctions seriously does not merely add "soft skills" to a technical curriculum. It cultivates the particular forms of human agency – accountable, relational, morally serious – that an AI-augmented economy will depend upon but cannot generate itself. These are precisely the competencies that constitute the third pillar of the AI Proficiency framework (organisational and human competencies), and it is no coincidence that they are also the ones that formal education has historically been most reluctant to measure.

Rethinking assessment is the operational expression of this ambition. If the final deliverable (an essay, a test score, or a design prototype) can now be generated by AI with minimal human contribution, then measuring the deliverable alone tells us very little about what a student has actually learned. Assessment in the Artificial Intelligence era must focus on transformation rather than output, in the same way that organisations must shift from measuring individual productivity to measuring the quality of human-AI synergy. Process-oriented evaluation instruments – including process logs, oral defences, Socratic seminars, and thinking-aloud protocols – measure not only what conclusion a student reaches but how their reasoning developed, where it changed, and why. In other words, what is needed is a focus on collaborative intelligence: rather than asking whether a student has produced a correct answer, educators must ask whether the student has become more capable through their engagement with AI, i.e., whether the scaffolding of machine assistance has supported genuine human growth or merely substituted for it. This is the educational equivalent of the IBR: not whether AI was used, but whether human engagement with it produced something that neither could have produced alone.

Conclusion: a roadmap for the coming years

Measuring human performance in the AI era requires a fundamental shift in the organizational and educational lens. The coming years will not be defined by the organizations and institutions that adopt AI most rapidly, but by those that most clearly understand what human beings uniquely contribute to AI-augmented systems and build the measurement, incentive, and educational architectures that develop and reward those contributions.

Three imperatives follow from the analysis presented here. First, organizations must redefine their metrics by implementing frameworks like AI Proficiency to assess technical, domain, and human competencies, and by adopting team-level indicators such as the HAR and the IBR. Measuring the quality of human-AI synergy, not merely the speed of output, is the foundational act from which all other changes follow.

Second, organizations must align their incentive structures with the capabilities they wish to develop. Rewarding AI proficiency in performance appraisals, promotion criteria, and peer recognition systems sends an organizational signal that the human-AI partnership is not a threat to career advancement but the primary vehicle for it.

Third, educational institutions must redesign their pedagogical and assessment models to cultivate the capabilities that genuinely matter in an intelligent economy. This means moving beyond the measurement of outputs to the assessment of cognitive transformation: prioritizing curiosity, judgment, ethical reflection, and the productive struggle that distinguishes learning from unreflective retrieval.

Taken together, these three imperatives define a coherent agenda: to ensure that the Artificial Intelligence era amplifies human potential rather than displacing it. The organizations and societies that act on this agenda will not only be more productive, but, even more importantly, they will be more genuinely human.

References

Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics (No. w24001). National Bureau of Economic Research.

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., & Lakhani, K. (2025). The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise (No. w33641). National Bureau of Economic Research

Microsoft (2025). Work Trend Index Annual Report

Olawade, D. B., Plabon, S. B., Ojo, A., Ogunbona, M. A., Makanjuola, B. D., & Olasilola, R. (2026). Human in the loop artificial intelligence in healthcare: applications, outcomes, and implementation challenges. *International Journal of Medical Informatics*, 106362.

Ponnusamy, A. A. N. (2025). Bridging LLM-generated code and requirements: reverse generation technique and SBC metric for developer insights. arXiv preprint arXiv:2502.07835.

Ranganathan, A., & Ye, X. M. (2026). AI Doesn't Reduce Work – It Intensifies It. *Harvard Business Review*.

Strauss, C., Strauss, E., & Kellermanns, F.W. (2026). Measuring AI Proficiency – A Multi-Dimensional Framework. Working Paper.

Strauss, E. (2025). Rethinking performance in the human-AI workplace, retrieved from https://escp.eu/thechoice/tomorrow-choices/rethinking-performance-in-the-human-ai-workplace/ on 27th of March 2026.