

# Intelligent organizational engineering driven by human-AI collaboration and explainable AI to increase productivity

Ingeniería organizacional inteligente impulsada por la colaboración entre humanos e IA y la IA explicable para aumentar la productividad

Francisco Herrera<sup>1</sup>

Recibido: 08/08/25 | Aceptado: 11/9/25

## Abstract

The rapid adoption of artificial intelligence (AI) in organizational engineering promises significant productivity gains, yet evidence shows a persistent “AI productivity paradox” in which task-level efficiencies fail to translate into measurable economic benefits. This paper examines how eXplainable AI (XAI) and human–AI collaboration (HAIC) can address this gap by aligning algorithmic capabilities with human expertise, trust, and organizational design. Drawing on recent empirical studies, we analyse the structural, cognitive, and socio-technical barriers that limit AI’s value realization, including inadequate integration, overreliance on automation, and inherent system opacity. We propose that XAI, embedded as both a technical and organizational capability, enables transparency, accountability, and adaptive collaboration across diverse stakeholder groups. Using a simulated organizational engineering scenario, we show how XAI-informed HAIC can enhance decision quality, redistribute cognitive workload, and foster iterative learning. The analysis underscores that AI’s real productivity potential lies not in automation alone, but in deliberate, human-centred integration that treats AI as a collaborative partner within resilient socio-technical systems driving intelligent organizational engineering to increase productivity.

**Keywords:** Organizational engineering, eXplainable artificial intelligence (XAI), human-AI collaboration (HAIC), productivity paradox, trust in AI.

## Resumen

La rápida adopción de la inteligencia artificial (IA) en la ingeniería de organización promete importantes aumentos de productividad, pero los estudios muestran una persistente «paradoja de la productividad de la IA», en la que la eficiencia a nivel de tareas no se traduce en beneficios económicos cuantificables. Este artículo examina cómo la colaboración entre humanos e IA (HAIC) y la IA explicable (XAI) pueden abordar esta brecha alineando las capacidades algorítmicas con la experiencia humana, la confianza y el diseño organizativo. Basándonos en estudios empíricos recientes, analizamos las barreras estructurales, cognitivas y sociotécnicas que limitan la realización del valor de la IA, incluyendo la integración inadecuada, la dependencia excesiva de la automatización y la opacidad inherente al sistema. Proponemos que la explicabilidad, integrada como capacidad tanto técnica como organizativa, permita la transparencia, la rendición de cuentas y la colaboración adaptativa entre los diversos grupos de interesados. Mediante un escenario simulado de ingeniería de organización, mostramos cómo la HAIC basada en XAI puede mejorar la calidad de las decisiones, redistribuir la carga de trabajo cognitiva y fomentar el aprendizaje iterativo. El análisis subraya que el verdadero potencial de productividad de la IA no reside únicamente en la automatización, sino en una integración deliberada y centrada en el ser humano que trata a la IA como un socio colaborador dentro de sistemas sociotécnicos resilientes que impulsan una ingeniería organizativa inteligente para aumentar la productividad.

**Palabras Clave:** Ingeniería organizacional, Inteligencia artificial explicable (XAI), colaboración entre humanos e IA (HAIC), paradoja de la productividad, confianza en la IA.

<sup>1</sup> Dept. of Computer Science and Artificial Intelligence, DaSCI Research Institute, University of Granada, 18071-Granada, Spain. E-mail: [herrera@decsai.ugr.es](mailto:herrera@decsai.ugr.es)

## 1. Introduction

Organizational engineering seeks to design, evaluate, and improve the structures, workflows, and decision-making systems that underpin modern institutions. In an era of accelerating technological advancement, particularly in artificial intelligence (AI), the foundations of productivity and innovation are being redefined. Gains once driven by human labour optimization and managerial efficiency are giving way to new frontiers—where intelligent systems augment cognitive work, automate decisions, and generate novel insights. Yet this transformation presents both vast opportunities and profound challenges.

AI is now indispensable across sectors, enabling predictive modelling, real-time optimization, and decision support in finance, manufacturing, healthcare, and human resources (Naudé et al., 2024). The rise of large language models like ChatGPT, Claude, and Gemini (Annepaka & Pakraym, 2025) and advances in reasoning approaches (Ke et al., 2025)—such as Wang et al.’s (2025a) three-level hierarchical reasoning—blur the traditional boundary between narrow AI and AGI, fuelling perceptions that certain AGI capabilities (Morris et al., 2024) may emerge soon. Alongside this, transformative AI (Gruetzemacher & Whittlestone, 2022; Lobo & Del Ser 2024), with its focus on societal-scale change, and agentic AI (Acharya et al., 2025), introducing autonomous goal-driven behaviour, are reshaping definitions of what is intelligent, impactful, or dangerous. These developments make clear that productive adoption cannot rely on raw computational power alone. Future readiness demands embedding explainability, trust calibration, and human–AI co-intelligence into organizational design, ensuring that transformative potential translates into sustainable, ethical productivity gains.

Yet adoption is not frictionless. As AI systems grow in complexity and opacity, concerns about trust, accountability, and oversight intensify. High-performing but inscrutable “black-box” models can erode confidence and stall integration into mission-critical operations, contributing to the AI productivity paradox (Chong, 2025): task-level efficiency gains often fail to yield proportionate organizational or economic benefits due to misalignment, underuse, or distrust.

To address this gap, eXplainable AI (XAI) has emerged as a cornerstone of responsible deployment. XAI encompasses models and design principles that make AI systems transparent and interpretable (Arrieta et al., 2020), framing explanation as a dynamic, context-sensitive dialogue between humans and algorithms (Herrera, 2025). Effective explanations must match users’ roles, goals, and decision contexts (Wang et al., 2025b), enabling trust calibration, accountability, and adaptive collaboration in organizational settings.

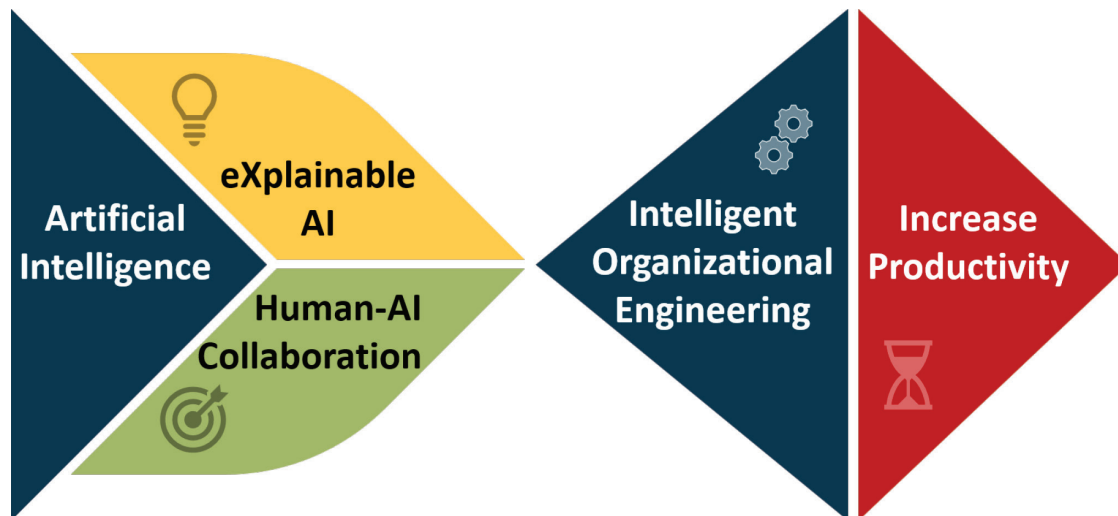
In parallel, human-AI collaboration (HAIC) is reshaping the landscape of organizational productivity. Unlike earlier models that focused on AI as a substitute for human labour, HAIC emphasizes synergy, co-creation, and hybrid intelligence. As Fragiadakis et al. (2024) propose, effective collaboration is built on shared goals, transparency, and mutual trust—where humans bring contextual awareness and ethical judgment, and AI offers speed, scale, and analytical rigor. This shift reflects a deeper socio-cognitive transformation in organizational engineering: moving from automation toward “*Co-intelligence*” (Mollick, 2024), where humans and machines learn from and with each other in real-time. Mehrotra et al. (2024) further highlight that the success of HAIC depends on cultivating appropriate trust, adaptivity, and user-centred design—especially in high-stakes or ambiguous decision environments.

This article addresses and emphasizes the hypothesis of how the convergence of XAI and HAIC can unlock new levels of productivity, resilience, and innovation within engineering organisations. It builds on the argument that technical performance alone is insufficient; instead, intelligent systems must be explainable, context-aware, and embedded within collaborative frameworks that reflect institutional goals and human capabilities. To frame this study, three interdependent challenges are examined, each essential to achieving productive, human-centred AI integration in engineering organizations.

- How can explainability facilitate better delegation and high-quality decision-making in human–AI teams?
- What collaborative models best align AI systems with expert human judgment to ensure synergy rather than substitution?
- What organizational engineering structures and cultural practices are required to support these hybrid systems through integration into workflows, governance, and institutional norms?

By bridging XAI and HAIC, this study contributes to a more nuanced understanding of how AI can augment—not replace—human expertise, enabling organizational engineering to achieve productivity that is not only measurable but also equitable, sustainable, and aligned with human values. Figure 1 shows the hypothesis emphasized in this essay, XAI and HAIC drive intelligent organizational engineering to increase productive.

The remainder of the paper is structured as follows: Section 2 outlines the productivity paradox of AI in modern organizations, offering macro- and micro-level perspectives on the mismatch between AI capabilities and realized value. Section 3 introduces XAI examining its role as a socio-technical enabler of trust and understanding. Section 4 develops an analysis of HAIC models, integrating concepts such as

**Figure 1.** XAI and HAIC driving intelligent organizational engineering to increase productivity.

co-intelligence and adaptive delegation to analyse productivity improvement. Section 5 presents a simulated case study of an engineering consultancy that implemented XAI and HAIC strategies to address design and project management inefficiencies. Section 6 discusses the broader implications for organizational design, risk, resistance, and trust in AI systems. Section 7 concludes the need on designing transparent, collaborative, and high-performance human-AI ecosystems.

## 2. The AI Productivity Paradox: Rethinking Value Realization in Organizations

Chong (2025) provides one of the most comprehensive analyses of the AI productivity paradox—the widening gap between task-level efficiency gains and meaningful economic benefits. Using large-scale empirical data, Chong shows that while 81% of office workers believe AI tools improve their performance and AI chatbots save 64–90% of time on certain tasks, only 3–7% of these gains translate into higher earnings. On average, 25 minutes saved per day—about 2.8% of work hours—produces negligible improvements in wages, job security, or organizational output.

A landmark Danish study tracking 25,000 workers across 7,000 AI-exposed workplaces reinforces this finding, reporting “precisely estimated zeros” in economic outcomes despite rapid adoption—jumping from 47% voluntary use to 83% with employer encouragement. The contrast with controlled experiments, which often show 15%+ productivity improvements, highlights the gap between laboratory results and real-world impact. Chong attributes this to structural barriers such as poor integration, inadequate training, unclear governance, and unrealistic organizational expectations. Echoing Gartner’s “trough of disillusionment,” he

argues that AI is often treated as a plug-in upgrade rather than a systemic transformation.

Acemoglu (2025) situates this paradox within a broader macroeconomic perspective, emphasizing that automation often produces “displacement effects” which can counteract efficiency gains. As AI substitutes certain job tasks, the net impact on labour demand and income distribution may be neutral or even negative unless new, complementary roles emerge. Without institutional reform or investment in human capital, AI-driven efficiency tends to concentrate value at the organizational or capital-owner level rather than distributing it across the workforce. Yet, beneath these macroeconomic challenges, a quieter shift is underway.

AI is generating entirely new categories of work—AI prompt designers, quality auditors, integration architects, and oversight coordinators—many of which fall under the emerging category of “AI shepherding.” Nearly 17% of AI users now perform such roles, the majority focusing on AI quality auditors, prompt engineers, and oversight coordinators. While these positions exemplify the human–AI complementarity potential described by both Chong and Acemoglu, the value they generate remains largely invisible to traditional productivity metrics or compensation systems.

The human toll is also significant. Many employees report burnout, escalating performance demands, and uncertainty about how to achieve promised AI gains. Kulesa et al. (2025) explore this challenge in the context of education, showing how AI tools—while enhancing personalization, reducing repetitive work, and providing rapid feedback—can also undermine the “productive struggle” that fosters deep learning, persistence, and critical thinking if used as a shortcut to answers.

Simkute et al. (2025) explain how poorly integrated AI can even reduce productivity, identifying “ironies” such as higher cognitive load from verifying outputs, workflow

disruptions, and skill erosion through overreliance. They argue for human-centred design, skill-based task allocation, and collaborative workflows to align AI capabilities with human expertise—reminding us that productivity comes not from inserting new tools into old systems but from rethinking socio-technical structures.

Organizations that actively support AI adoption—through training, custom tool development, and ethical oversight—report narrower gender and age gaps in use, higher employee satisfaction, and, in some cases, modest wage gains. Whittle (2025) cautions, however, that time savings and efficiency improvements at the micro level rarely scale into macro-level growth. In some cases, verification, coordination, and mitigation of overreliance can increase workloads. Whittle challenges the assumption that faster task completion equals higher productivity, suggesting that AI’s transformative potential may instead lie in enabling deeper reflection, creativity, and innovation.

Ultimately, the productivity paradox is less a failure of AI than of deployment, organizational engineering, and measurement. Real gains require leadership commitment, institutional design, human-centred integration, and XAI and HAIC that foster trust, understanding, and the ability to shape outputs in context. Organizations that treat AI as a collaborative partner—rather than a replacement—will be best positioned to thrive in an augmented future.

### 3. Explainable AI as a Foundation for Organizational Intelligence

XAI comprises a diverse set of models, methodologies, and design principles aimed at making the internal workings and outputs of AI systems explainability and meaningful to human users (Arrieta et al., 2020; Herrera, 2025). As AI becomes a staple in high-risk domains—such as finance, healthcare, and organizational engineering—the challenge is no longer solely about optimizing predictive accuracy. Instead, organizations are tasked with ensuring that AI systems are transparent, accountable, and responsive to the needs and expectations of human collaborators. XAI emerges as a key mechanism to bridge the cognitive divide between opaque “black-box” models and domain experts, decision-makers, and stakeholders who must understand and trust the AI systems they rely on.

More than just a technical toolset, XAI represents a socio-technical paradigm. According to (Herrera, 2025), explanation in AI is not a static output but a dynamic, communicative process—one that involves interpretation, adaptation, and responsiveness to context. XAI, therefore, should not be treated as an afterthought or compliance checkbox; rather, it must be embedded into the very architecture of human-AI workflows. This stakeholder-centred view introduces a critical distinction between algorithmic

transparency (how the AI functions internally) and explanatory usefulness (how that understanding is conveyed meaningfully to users). Herrera (2025) identifies six distinct stakeholder groups whose explanatory needs vary substantially. Developers seek detailed debugging and technical insights to refine models; managers require actionable narratives that link AI outputs to strategic and operational goals; auditors demand procedural accountability to ensure compliance with legal and organizational standards; affected parties need justifications that help them understand or contest decisions impacting them; policy-makers require clear explanations of system limitations, risks, and societal implications to inform regulation; and the general public looks for accessible, non-technical explanations that foster transparency and trust in AI systems.

Recognizing these varying epistemic needs, XAI becomes a form of organizational attentiveness—an institution’s ability to listen, adapt, and respond to how people engage with AI systems across roles and power dynamics. For example, in engineering organizations, managers may need streamlined visualizations of AI-based performance assessments, while design engineers may require detailed justifications for predictive failure modes. Thus, explanations must not only clarify algorithmic logic but also align with professional norms, regulatory expectations, and human interpretive capacity.

Wang et al. (2025b) underscore this point, arguing that “good” XAI design is inherently contextual: the same explanation may be helpful to one stakeholder and misleading to another. Their study illustrates that effective explanation interfaces must be sensitive to user goals, decision stakes, and cognitive resources. This implies that engineering organizations adopting XAI should consider role-adaptive interfaces and modular explanation strategies tailored to varied workflows and decision contexts.

Practically, XAI encompasses a range of techniques—saliency maps, rule-based models, SHAP/LIME interpretability layers, counterfactuals, and interactive explanations. Yet Herrera (2025) cautions that over-reliance on technical mechanisms without explainability training or alignment with user practices can lead to “explanation fatigue” or misinterpretation. Hence, XAI must be integrated into broader systems of user education, governance, and iterative co-design to avoid overreliance.

In organizational engineering, XAI can play three pivotal roles:

- Calibrating trust, avoiding both blind reliance and unfounded scepticism.
- Enabling accountability, clarifying who is responsible for what decisions.
- Supporting adaptive collaboration, allowing humans and AI to co-adjust their behaviour based on shared

understanding. It transforms AI from a passive tool into an active partner in decision-making.

The productivity gains from XAI stem directly from this understanding. When experts understand why AI systems make certain recommendations, they are more capable of identifying edge cases, refining input parameters, and accelerating decision cycles. Instead of second-guessing opaque suggestions, engineers and managers can engage in constructive scrutiny—reframing decisions, adapting processes, or improving data inputs. This cycle of human-AI co-learning not only boosts individual decision accuracy but enhances organizational learning over time.

Moreover, explainability fosters systemic agility. Organizations can more readily identify mismatches between model behaviour and business logic, fine-tune delegation schemes, and develop feedback loops that continually align AI functionality with strategic goals. This feedback-rich environment supports both short-term task optimization and long-term transformation.

To finish this analysis of XAI, we must also point out the opacity inherent to any AI systems. Opacity in AI emerges not merely as a technical hurdle but as an enduring, multifaceted condition that demands governance rather than eradication. As Hähnel and Hauswald (2025) note, opacity—whether stemming from inscrutable model internals, data abstraction, or institutional secrecy—can undermine trust, yet total transparency is neither always possible nor sufficient for legitimacy. Herrera and Calderón (2025) propose the LoBOX (*Lack of Belief: Opacity and eXplainability*) framework, making this explicit by distinguishing *accidental opacity*, which can be reduced through better design and communication, from *per se opacity*, intrinsic to complex models and resistant to full interpretability. In such cases, institutional trust becomes the anchor: layered, role-sensitive explanations; procedural accountability; and independent oversight shift the trust burden from individual comprehension to systemic credibility. Freiman et al. (2025) extend this point by showing that trust and opacity must be understood in technical, institutional, and epistemic dimensions, with governance structures calibrated to context. Together, these perspectives argue that fixing XAI's limitations is less about chasing exhaustive transparency than about embedding opacity within ethically justifiable, institutionally grounded frameworks that enable scrutiny, contestation, and context-appropriate intelligibility.

In sum, XAI should be understood as a cornerstone of human-AI synergy—an architectural and cultural principle that aligns intelligent systems with the socio-technical fabric of organizations. Its contribution to productivity lies not only in making systems understandable but also in making them usable, improvable, and accountable across an evolving landscape of roles, risks, and responsibilities.

This includes acknowledging and managing the inherent opacity of AI systems—distinguishing what can be clarified through design and communication from what is intrinsically resistant to full interpretability. By embedding such opacity within governance frameworks that ensure procedural accountability, role-sensitive explanations, and institutional trust, organizations can sustain both confidence and critical oversight in HAIC.

#### 4. Human-AI Collaboration Models: Toward Augmented Decision-Making

HAIC represents a significant evolution in how organizations interact with intelligent systems, moving from traditional automation paradigms toward collaborative, adaptive, and human-centred frameworks. Unlike automation, which typically displaces human labour, HAIC emphasizes the augmentation of human capabilities through partnership with AI agents. This collaboration integrates the strengths of both parties—humans bring ethical reasoning, contextual judgment, and creativity, while AI contributes speed, scalability, and pattern recognition. As Fragiadakis et al. (2024) propose, effective HAIC requires a dynamic and bidirectional interaction model that incorporates shared goals, transparent communication, and continuous feedback. Their methodological framework defines three core modes of HAIC—human-centric, AI-centric, and symbiotic—each demanding different configurations of task allocation and trust calibration. These models reflect the importance of aligning AI system design with human cognition, values, and workflows to realize gains in both performance and productivity.

In the context of Industry 5.0, the significance of HAIC extends beyond operational efficiency to become a strategic imperative. Krause et al. (2024) argue that the next wave of industrial transformation emphasizes human-centricity, resilience, and sustainability, shifting the role of AI from mere automation to meaningful augmentation. Their study highlights the role of knowledge graphs in contextualizing human-AI interactions, enabling machines to understand complex relationships between tasks, roles, and organizational goals. However, the authors also warn of persistent challenges—particularly around explainability, trust, and governance—that must be addressed to unlock the full potential of collaborative intelligence. In this paradigm, AI systems should be designed not only to optimize performance but to empower human workers, support inclusive decision-making, and foster organizational learning.

A growing body of empirical research supports the productivity-enhancing potential of HAIC when combined with XAI. Senoner et al. (2024) conducted a controlled study showing that users working with AI systems that

provided transparent explanations consistently outperformed those without such support. The presence of visual and textual justifications improved task accuracy, boosted user confidence, and fostered a stronger sense of control. These benefits were particularly pronounced in scenarios involving complex decisions or high uncertainty, suggesting that explainability functions as a cognitive scaffold. Rather than passively observing AI outputs, users engaged more critically and constructively, resulting in a more effective division of cognitive labour between human and machine.

Trust and delegation are central to the success of HAIC. Wen et al. (2025) demonstrate that perceived AI trustworthiness directly influences the degree to which human users integrate AI recommendations into decision-making. Their experiments reveal that excessive autonomy in AI systems—especially when they appear too agentic—can reduce user trust and diminish collaborative effectiveness. This aligns with the Socio-Cognitive Model of Trust, which posits that trust evolves through recognition, understanding, and appropriate delegation. Wen et al.'s findings underscore the importance of role-sensitive design: AI systems must communicate their logic and limitations clearly, positioning themselves as reliable partners rather than opaque authorities.

Similarly, Hemmer et al. (2023) explore the impact of AI delegation on human performance and satisfaction. Their study identifies a critical balance: while moderate AI delegation enhances both accuracy and user satisfaction, extremes on either side—total human control or full AI autonomy—lead to suboptimal outcomes. Users reported higher engagement and performance when the AI system offered support without undermining human agency. This highlights the value of calibrated delegation mechanisms and adaptive interfaces that respond to user expertise, task complexity, and situational context. Explainability again emerges as a key enabler of this balance, ensuring that users understand not just what the AI recommends, but why.

At the heart of these findings lies the concept of co-intelligence—a paradigm that transcends simple augmentation by fostering deep, mutual co-creation between human and AI agents. Mollick (2024) characterizes co-intelligence as a collaborative process where both actors iterate, adapt, and generate novel solutions together. Unlike traditional models focused on efficiency or automation, co-intelligence prioritizes shared learning, exploration, and creativity. In the context of engineering organizations, this approach can be considered as especially transformative. Engineering work is inherently interdisciplinary and innovation-driven; embedding co-intelligent systems into these workflows enables teams to explore design alternatives, validate models, and simulate strategies with unprecedented agility and depth.

This vision is reinforced by Alam et al. (2024), who argue that next-generation AI systems should be designed to extend—not replace—engineering expertise. Their MIT study outlines how HAIC in design and manufacturing enables engineers to evaluate trade-offs, uncover hidden constraints, and iterate more effectively. These tools, when explainable and co-creative, allow organizations to align AI capabilities with human goals, ultimately driving gains in quality, adaptability, and productivity.

The transition from automation to augmentation, and from augmentation to co-intelligence, represents a strategic frontier for organizational engineering. Embracing this shift will be essential for organizations seeking not just efficiency, but long-term innovation and resilience.

## 5. Case Study: Embedding XAI and HAIC in Engineering Workflows

To illustrate the integration of XAI and HAIC in organizational engineering, we introduce a simulated case study involving a mid-sized engineering consultancy specializing in infrastructure design and maintenance. The firm faced persistent productivity challenges due to delays in project planning, inconsistent evaluation of design alternatives, and frequent misalignment between technical and managerial teams. In response, the organization adopted a dual-system AI integration strategy: a predictive analytics engine to assist in engineering simulations, and a performance evaluation tool for project management—both enhanced with XAI interfaces.

**Scenario Design.** The AI components were designed to operate as co-intelligent agents within cross-functional teams. The predictive engine provided design recommendations (e.g., load-bearing calculations, cost optimizations, and compliance forecasts), while the performance evaluator offered managerial insights into team efficiency, resource allocation, and project timelines. Each system incorporated explainability layers including counterfactual analysis, rule-based justifications, and uncertainty visualizations. Moreover, role-adaptive explanation panels were introduced—technical staff received granular model diagnostics, while managers accessed high-level impact summaries and what-if simulations. HAIC protocols were established to support adaptive delegation: team leads could either accept AI recommendations directly, revise inputs based on AI feedback, or override suggestions with contextual justifications.

**Observed Impacts.** After three months of implementation, several key results emerged. First, the engineering teams reported a 25% reduction in the time spent on design iteration cycles, attributed to improved understanding of AI-generated trade-offs. Rather than treating AI suggestions as opaque directives, designers engaged in back-and-forth refinement with the system, accelerating

convergence on optimal solutions. Second, the firm experienced a 15% improvement in project schedule adherence, largely due to better forecasting of resource constraints and clearer communication between managerial and technical functions.

Qualitative feedback further indicated increased trust and perceived competence of AI systems, particularly when explanations were provided in real-time. Designers cited examples where explanations of compliance failures in structural simulations helped them discover overlooked constraints, while project managers used performance rationales to reassign tasks more fairly and effectively.

Importantly, this collaborative success hinged on transparency. Employees emphasized that without explainable output, they would have resisted adopting the tools. By embedding XAI features and aligning explanations with stakeholder needs.

**Organizational Learning and Co-Creation.** Beyond performance metrics, the implementation catalyzed a shift toward co-intelligence. Weekly team retrospectives were used not only to evaluate project progress but also to co-evolve AI models. Domain experts provided feedback on erroneous predictions, which was used to update the AI's feature weighting or retrain components. In effect, the AI systems became adaptive partners that learned from human interaction, while human workers enhanced their own decision-making by engaging critically with machine outputs. This iterative loop of mutual adaptation and co-creation exemplifies the transformation from automation to augmentation, and finally, to co-intelligence.

The integration of explainability and collaboration is transforming engineering productivity, not solely by chasing raw efficiency gains, but by enabling smarter, faster, and more accountable workflows. By combining human expertise with algorithmic insights, organizations can evolve from simple automation toward true augmentation, where decision-making is both more informed and more transparent (Alam et al., 2024).

## 6. Discussion: Strategic, Ethical, and Structural Implications

The rise of transformative AI—systems capable of reshaping entire sectors through large-scale automation, advanced reasoning, and adaptive learning—forces organizational engineering to move beyond incremental process improvements toward deliberate future readiness. For engineering organizations, this means embedding AI within structural designs that are resilient to technological change, adaptable to shifting stakeholder needs, and grounded in ethical and operational transparency. Future readiness is not simply about acquiring powerful AI tools; it requires translating capabilities into design principles such as role-adaptive

workflows, human-in-the-loop oversight, and continuous feedback systems that recalibrate decision-making processes in real time. By treating transformative AI as both a technological and organizational capability, leaders can ensure that adoption amplifies human expertise, supports strategic objectives, and avoids the productivity traps of poorly integrated innovation—laying the groundwork for the strategic AI integration models discussed in this section.

The discussion that follows draws from recent empirical studies and conceptual frameworks to examine both the strategic implications and inherent risks associated with AI adoption in engineering organizations. It emphasizes that realizing the full potential of AI—particularly generative and explainable systems—requires aligning innovation with human agency, ethical oversight, and institutional resilience. Section 6 thus serves as a bridge between theory and practice, showing how engineering organizations can turn AI's transformative potential into ethically grounded, human-centred, and strategically aligned innovation.

### 6.1. Organizational Design for Strategic AI Integration

Effective AI adoption in engineering organizations is not simply a matter of deploying advanced tools—it requires carefully structuring the organizational environment in which those tools operate. Strategic AI integration demands aligning technology capabilities with human workflows, cultural norms, and decision-making processes so that automation complements, rather than disrupts, human expertise.

The work of Gafni et al. (2024) offers compelling evidence that AI systems can be designed to enhance objectivity in domains typically dominated by human subjectivity, such as soft skills evaluation. By embedding AI into human resource processes, organizations can mitigate human biases and foster fairer assessments of competencies like communication, teamwork, and adaptability—qualities that are increasingly vital in cross-functional engineering environments. However, their study also highlights a key organizational insight: the success of AI integration hinges on user acceptance and perceived fairness. Transparent explainability, contextualized feedback, and shared control with human evaluators are essential to ensure trust and avoid resistance.

As Holmström and Carroll (2025) emphasize, innovation with generative AI requires more than tool adoption—it demands strategic alignment with organizational goals, capabilities, and governance structures. Their framework shows that successful AI-driven innovation emerges when generative technologies are embedded into workflows that amplify human creativity, improve decision-making, and unlock new value creation. For engineering organizations, this means designing collaborative systems where generative AI accelerates iteration and design while reinforcing

accountability and human oversight, positioning AI as an organizational capability rather than a mere technical upgrade. To support this, Holmström and Carroll (2024) propose a strategic typology that helps managers balance automation and augmentation:

- Traditional Tool (low automation, low augmentation)
- Basic Automation (high automation, low augmentation)
- Automated Assistance (low automation, high augmentation)
- Assisted Augmentation (high automation, high augmentation)

This framework helps managers align AI usage with innovation goals—whether it’s streamlining processes or co-creating novel solutions. It underlines the idea that AI should augment human creativity, not replace it, matching well with your advocacy for co-intelligence and XAI-enabled collaboration.

## 6.2. Resilient Human–AI Systems: Managing Risk, Resistance, and Trust in Organizational Engineering

While the integration of AI systems into organizational workflows holds great promise for enhancing productivity, collaboration, and decision-making, it also introduces a constellation of risks that, if unmanaged, can erode the very benefits these technologies aim to deliver. Among the most pressing concerns is over trust, a phenomenon in which users rely too heavily on AI outputs without adequate critical reflection. As Mehrotra et al. (2024) argue, trust in AI must be calibrated, not maximized. Excessive reliance, especially in the absence of robust explainability, can lead users to defer judgment, ignore edge cases, or accept flawed recommendations, ultimately undermining organizational resilience and decision quality.

Closely linked to this is the erosion of human skills through sustained cognitive offloading. As AI systems increasingly take on analytical, evaluative, and even creative tasks, there is a risk that human collaborators may lose proficiency in essential domains, including critical thinking, intuition, and contextual judgment. This issue is especially acute in engineering organizations, where tacit knowledge and domain expertise are central to innovation. Holmström and Carroll (2025) caution that while generative AI tools can streamline ideation and simulation processes, they must be framed as augmentation tools—not replacements—for human ingenuity. Otherwise, organizations may experience a slow decay of internal expertise, reducing long-term adaptability and strategic flexibility.

Beyond individual or task-level concerns, AI adoption also presents organizational and structural risks. Employee resistance often emerges when AI is perceived as

a mechanism for surveillance, control, or workforce reduction. This psychological pushback can reduce adoption rates, fuel distrust, and diminish the collaborative potential of AI systems. Addressing this requires not only technical safeguards but also inclusive, participatory governance frameworks that define accountability, preserve user agency, and promote ethical AI use. Training programs, co-design practices, and ongoing feedback loops can empower employees to engage meaningfully with AI, reducing friction and reinforcing human-centred innovation cultures.

Furthermore, the work of Abercrombie et al. (2024) provides a crucial lens to understand AI risks as systemic and multi-layered. Their taxonomy shows that AI harms go beyond technical errors and include broader societal concerns such as automation bias, erosion of responsibility, and structural discrimination. These are not merely design flaws—they are embedded in how AI systems are implemented and governed. Complementing this, Zhang et al. (2025) expose the emergent risks of prolonged human-AI interaction, including emotional dependence, misaligned goals, and inflation of AI authority. Together, these perspectives emphasize that HAIC must be proactively shaped, not passively accepted.

At the heart of managing these risks lies the dynamic and fragile construct of trust. As Afroogh et al. (2024) emphasize in their comprehensive review, trust in AI is not monolithic—it must be contextual, multidimensional, and continuously monitored. It spans technical competence, reliability, ethical alignment, and user comprehension. Importantly, misplaced or blind trust can be just as harmful as distrust, leading to automation complacency and ethical blind spots. Afroogh et al. (2024) underscore the need for user-centred design, transparency, and ongoing communication to foster appropriate trust levels that evolve alongside user experience and task demands. This perspective directly reinforces the argument for XAI and co-intelligent design strategies in this paper, as they help maintain human oversight, support informed delegation, and sustain long-term collaborative efficacy.

The future of AI trustworthiness will be built not on blind adoption, but on a deliberate and reflective co-evolution between humans and machines. In this paradigm, explainability, context-awareness, and ethical design form the cornerstones of resilient organizational intelligence, ensuring that AI systems enhance rather than erode human judgment.

Revisiting the three challenges outlined in the introduction, our analysis of strategic AI integration shows:

- First, that explainability remains the linchpin for effective delegation and decision quality in human–AI teams. Without transparent, context-aware explanations, trust calibration is fragile, and decision-making defaults either to blind reliance or rejection.

- Second, the collaborative models that most effectively align AI systems with expert human judgment are those that embed role-sensitive interfaces and maintain human-in-the-loop oversight, ensuring synergy rather than substitution.
- Third, sustaining these hybrid systems over time demands organizational engineering structures and cultural practices that normalize co-intelligence — integrating AI capabilities into workflows, governance mechanisms, and institutional norms so they remain adaptive to technological change and stakeholder needs.

Together, these elements form a practical blueprint for moving from technical capability to productive, human-centred AI adoption—advancing the vision of intelligent organizational engineering. This approach reflects the essence of resilience in human–AI systems: aligning innovation with trust, adaptability, and sustained human oversight.

In sum, while the promise of AI in organizational engineering is significant, its success depends on our ability to balance innovation with reflection, and automation with accountability. Addressing the risks of over trust, skill erosion, systemic harm, and resistance requires a holistic, socio-technical strategy—where human values, organizational design, and AI capabilities evolve together toward more resilient, ethical, and productive futures.

## 7. Conclusion: Designing Transparent and Co-Intelligent Organizations

As AI systems become increasingly integrated into organizational engineering, the challenge is no longer whether to adopt AI, but how to do so in ways that are effective, ethical, and productive. This paper has explored the critical intersection of XAI and HAIC driving intelligent organizational engineering as a strategic foundation for addressing the productivity paradox—that is, the persistent gap between technological capabilities and realized organizational value.

As highlighted in Section 2, the productivity paradox reveals that while AI tools are widely adopted and offer measurable time-saving benefits, these gains often fail to translate into meaningful economic outcomes or sustained organizational performance. This disconnect underscores the need for more than technological implementation—it calls for human-centred integration. XAI and HAIC, when deployed as co-enablers, offer a path to bridging this gap by enhancing human understanding, aligning AI recommendations with expert judgment, and fostering more agile and accountable decision ecosystems. In short, solving the productivity paradox requires reframing AI not as a replacement force, but as a partner in collaborative transformation.

Our analysis emphasizes that explainability is not a peripheral technical feature but a core enabler of trust,

transparency, and collaborative adaptability. When explanations are tailored to the cognitive and contextual needs of stakeholders—developers, decision-makers, or affected users—they support not only legal and organizational accountability but also foster informed human oversight. In this way, XAI serves as the interpretive bridge that renders AI systems usable, auditable, and aligned with institutional priorities.

Simultaneously, HAIC represents a paradigm shift from automation to co-intelligence. Rather than displacing human agency, HAIC emphasizes synergistic task-sharing, dynamic delegation, and contextualized support. Empirical research cited throughout this work confirms that human-AI partnerships—when grounded in mutual trust and calibrated responsibility—outperform purely human or fully autonomous systems in terms of productivity, decision quality, and user satisfaction.

For industrial and organizational engineering contexts, the implications are clear: unlocking AI’s potential demands investment in role-adaptive explanation systems, ethically aware delegation frameworks, and human-centred governance. These capabilities, when embedded into engineering workflows, foster *intelligent productivity*, measured not just in efficiency gains, but in resilience, adaptability, and innovation capacity. Trust in AI must be actively cultivated—not passively assumed.

The future of productive organizations depends not on how quickly they adopt AI, but on how wisely they co-evolve with it.

## Funding

This publication is part of the project “Ethical, Responsible, and General Purpose Artificial Intelligence: Applications In Risk Scenarios” (IAFER) Exp.:TSI-100927-2023-1 funded through the creation of university-industry research programs (Enia Programs), aimed at the research and development of artificial intelligence, for its dissemination and education within the framework of the Recovery, Transformation and Resilience Plan from the European Union Next Generation EU through the Ministry of Digital Transformation and the Civil Service. This work was also partially supported by Knowledge Generation Projects, funded by the Spanish Ministry of Science, Innovation, and Universities of Spain under the project PID2023-150070NB-I00.

## Declaration of AI-assisted technologies in the writing process

During the preparation of this work, the author used large-language models to improve the readability and language of the manuscript. After using this tool/service,

the author reviewed and edited the content as needed and assumed full responsibility for the content of the published article.

## References

- ABERCROMBIE, G., et al. (2024). A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*.
- ACEMOGLU, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13-58.
- ACHARYA, D. B., et al. (2025). Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13, 18912-18936.
- AFROOGH, S., et al. (2024). Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1), 1-30.
- ALAM, M. F., et al. (2024) From Automation to Augmentation: Redefining Engineering Design and Manufacturing in the Age of NextGen-AI. *An MIT Exploration of Generative AI*, March. <https://doi.org/10.21428/e4baedd9.e39b392d>.
- ANNEPAKA, Y., & PAKRAY, P. (2025). Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems*, 67(3), 2967-3022.
- ARRIETA, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- CHONG, N. S. T. (2025). The AI Productivity Paradox: Why Your AI-Powered Workday Isn't Making You Richer. UN Blog. <https://c3.unu.edu/blog/the-ai-productivity-paradox-why-your-ai-powered-workday-isnt-making-you-richer>. Access July 9, 2025.
- FRAGIADAKIS, G., et al. (2024). Evaluating human-AI collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*.
- FREIMAN, O., et al. (2025). 'Opacity' and 'Trust': From Concepts and Measurements to Public Policy. *Philosophy & Technology*, 38(1), 29.
- GAFNI, et al. (2024). Objectivity by design: The impact of AI-driven approach on employees' soft skills evaluation. *Information and software technology*, 170, 107430.
- GRUETZEMACHER, R., & WHITTLESTONE, J. (2022). The transformative potential of artificial intelligence. *Futures*, 135, 102884.
- HÄHNEL, M., HAUSWALD, R. (2025). Trust and Opacity in Artificial Intelligence: Mapping the Discourse. *Philosophy & Technology*, 38(3), 115.
- HEMMER, P., et al. (2023). Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 453-463).
- HERRERA, F. (2025). Reflections and attentiveness on eXplainable Artificial Intelligence (XAI). The journey ahead from criticisms to human-AI collaboration. *Information Fusion*, 121, 103133.
- HERRERA, F., & CALDERÓN, R. (2025). Opacity as a Feature, Not a Flaw: The LoBOX Governance Ethic for Role-Sensitive Explainability and Institutional Trust in AI. *arXiv preprint arXiv:2505.20304*.
- HOLMSTGRÖM, J., & CARROLL, N. (2025). How organizations can innovate with generative AI. *Business Horizons*, 68 (5), 559-573.
- KE, Z., et al. (2025). A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.
- KRAUSE, F., et al (2024). Managing human-AI collaborations within industry 5.0 scenarios via knowledge graphs: key challenges and lessons learned. *Frontiers in Artificial Intelligence*, 7, 1247712.
- KULESA, A. C., et al. (2025). Productive Struggle: How Artificial Intelligence Is Changing Learning, Effort, and Youth Development in Education. *Bellwether*.
- LOBO J. L., & DEL SER, J. (2024). Can transformative AI shape a new age for our civilization? Navigating between speculation and reality. *arXiv preprint arXiv:2412.08273*.
- MEHROTRA, S., et al. (2024). A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4), 1-45.
- MOLLICK, E. (2024). Co-intelligence: Living and working with AI. *Penguin*.
- MORRIS, M. R., et al. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*.
- NAUDÉ W., et al. (2024). Artificial Intelligence: Economic Perspectives and Models. *Cambridge University Press*.
- SENONER, J., et al. (2024). Explainable AI improves task performance in human-AI collaboration. *Scientific reports*, 14(1), 31150.
- SIMKUTE, A., et al (2025). Ironies of generative AI: understanding and mitigating productivity loss in Human-AI interaction. *International Journal of Human-Computer Interaction*, 41(5), 2898-2919.
- WANG, G., et al. (2025a). Hierarchical Reasoning Model. *arXiv preprint arXiv:2506.21734*.
- WANG, L., et al. (2025b). " Good" XAI Design: For What? In Which Ways? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-13).
- WEN, Y. et al. (2025). Trust and AI weight: human-AI collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, 3, 1419403.
- WHITLE, J. (2025) Does AI actually boost productivity? The evidence is murky. *The conversation*. <https://theconversation.com/does-ai-actually-boost-productivity-the-evidence-is-murky-260690>
- ZHANG, R., et al. (2025). The dark side of AI companionship: A taxonomy of harmful algorithmic behaviors in human-AI relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).