Reversing the Paradigm: Building AI-First Systems with Human Guidance

Cosimo Spera[1*] and Garima Agrawal[1†]

[1*]Minerva CQ, USA.

*Corresponding author(s). E-mail(s): cosimo@minervacq.com;
Contributing authors: garima@minervacq.com;
†These authors contributed equally to this work.

Abstract

The relationship between humans and artificial intelligence is no longer a concept from science fiction—it's a growing reality reshaping how we live and work. Today, AI isn't just confined to research labs or tech giants' headquarters; it's embedded in everyday experiences, quietly powering customer service chats, personalizing travel recommendations, assisting doctors with diagnoses, and supporting educators in classrooms.

What makes this moment in AI's evolution particularly compelling is its increasing collaborative nature. Rather than replacing humans, AI is augmenting our capabilities—streamlining routine tasks, enhancing decision-making with data-driven insights, and fueling creativity in fields like design, music, and writing. The future of work is shifting toward AI agents taking the lead on tasks, with humans stepping in as supervisors, strategists, and ethical stewards—flipping the traditional model on its head. Instead of humans driving every action with AI as a support tool, intelligent agents will increasingly operate autonomously within defined parameters, handling everything from scheduling and customer interactions to complex decision-making workflows. Humans will guide, monitor, and fine-tune these agents to ensure alignment with organizational goals, values, and context. This shift promises significant gains in efficiency and scalability but also demands a new mindset—one that prioritizes meaningful collaboration between autonomous AI systems and thoughtful human guidance.

As AI agents assume more autonomous roles, the potential benefits are substantial: increased productivity, accelerated decision-making, cost savings, and the ability to scale operations in unprecedented ways. However, these benefits come with notable risks—including reduced human control, algorithmic bias, security vulnerabilities, and a widening skills gap. To navigate this transition successfully, organizations and societies must rethink roles, invest in upskilling, embed ethical frameworks into AI development, and prioritize transparency and accountability. This paper explores the technological and organizational shifts required to responsibly transition to AI-first systems—where autonomy is balanced with human values, guidance, and strategic intent.

Keywords: Artificial Intelligence, AI-First Systems, AI Agents, Agentic AI, Human-in-the-Loop, Human-Guided AI Systems, AI-Human Collaboration, Responsible AI, Autonomous Systems

Paper Organization

This paper begins with an overview of the current AI landscape, followed by a discussion of the prevailing paradigm: "Humans in the driver's seat, with AI in a supporting role." The third section presents the case for reversing this paradigm—advocating for AI-first systems where humans provide guidance rather than direct control. The fourth and fifth sections explore the design principles for such systems and examine associated risks. The sixth section highlights current AI-first use cases, while the seventh outlines a timeline for transitioning toward this model. The paper concludes with reflections on designing responsible, scalable AI-first systems guided by human values.

1 Introduction

The emergence of generative AI and autonomous systems marks a significant technological inflection point, with far-reaching implications across domains. Generative AI models—such as large language models and generative adversarial networks—demonstrate unprecedented capabilities in producing coherent text, realistic images, and functional code, effectively automating aspects of creativity and problem-solving. Concurrently, autonomous systems—ranging from robotics and drones to self-driving vehicles—are becoming increasingly adept at perceiving their environments, making decisions, and executing actions without direct human involvement. These advances promise increased efficiency and innovation but also raise challenges related to safety, accountability, and the socio-economic consequences of automation.

The current AI paradigm primarily positions artificial intelligence as a tool or assistant to human operators. In this model, AI augments human capabilities—supporting decision-making, automating routine tasks, and extracting insights from data—while humans retain primary control and responsibility. This *human-in-the-loop* approach emphasizes interpretability, transparency, and ethical safeguards, ensuring that AI operates within boundaries defined by human judgment. The assistant paradigm has enabled productivity gains across sectors such as healthcare, finance, education, and customer service, while mitigating risks associated with full autonomy.

An emerging shift in AI research and deployment envisions a reversal of this traditional human-centric model. In this evolving paradigm, AI systems act as the primary operators, with humans stepping into guiding, supervisory, and ethical roles.

This inversion recognizes the increasing capability of autonomous AI systems to manage complex, high-frequency decisions with speed, scalability, and consistency. In such systems, humans no longer drive every action but instead provide contextual judgment, strategic input, and system-level intervention when necessary. This transformation in roles has profound implications for how we design, govern, and interact with AI systems—demanding a new framework for trust calibration, accountability, and collaboration in AI-first environments.

We define *AI-first systems* as socio-technical architectures in which AI agents serve as the primary operational entities—autonomously performing tasks, making decisions, and coordinating workflows—while humans adopt supervisory, strategic, or ethical guidance roles. Unlike traditional human-in-the-loop models, where AI augments human decision-making, AI-first systems invert this dynamic by embedding autonomy into the core execution layer, with humans guiding and refining system-level behavior as needed. This design shift enables greater scalability, responsiveness, and efficiency in domains where algorithmic decision-making can outperform human-led processes.

2 The Current Paradigm: Humans in the Driver's Seat

AI technologies have become deeply integrated into everyday digital experiences through use cases such as copilots, chatbots, and recommendation engines. Copilot systems—exemplified by tools like GitHub Copilot or AI writing assistants—augment human productivity by generating context-aware suggestions in real time, streamlining tasks in software development, writing, and data analysis. Chatbots, powered by natural language processing, serve as front-line interfaces in customer service, healthcare triage, and education, offering scalable, round-the-clock engagement while reducing the burden on human agents. Recommendation engines, widely deployed in e-commerce, entertainment, and social media platforms, personalize user experiences by predicting preferences based on behavioral data, thereby increasing engagement and optimizing content delivery. Collectively, these applications illustrate how AI functions as an assistive layer—enhancing efficiency, personalization, and accessibility across a range of domains.

Positioning AI as an assistant, with humans in the lead decision-making role, offers several practical advantages—particularly in terms of control, trust, and legal clarity. First, it ensures that human operators maintain ultimate authority over key decisions, with AI confined to supporting functions within well-defined boundaries. Second, trust is reinforced when AI behavior remains interpretable, predictable, and subject to human validation, allowing for real-time corrections and calibration. This structure also simplifies accountability: when humans are explicitly in the driver's seat, it becomes easier to assign responsibility, ensure regulatory compliance, and enforce ethical standards. These characteristics have made the current paradigm well-suited to environments where transparency, caution, and institutional trust are paramount.

However, this model is not without limitations—particularly as AI capabilities continue to scale. Human guidance in high-stakes or ambiguous scenarios can become a bottleneck, constraining the scalability and responsiveness of AI systems. Latency may increase when decisions are deferred to human input, especially in time-sensitive contexts. Supervisors may experience cognitive overload when required to monitor multiple AI agents or interpret complex outputs without adequate abstraction and summarization tools. Furthermore, the continuous operation of advanced AI models raises concerns about energy consumption and long-term sustainability. Finally, deploying and maintaining such hybrid systems requires significant investment in infrastructure, personnel training, and human-in-the-loop frameworks—factors that may limit widespread adoption across industries.

3 The Case for Reversal: Why and When AI Should Lead

AI should be positioned as the lead operator in socio-technical systems when the capabilities of artificial agents surpass human limitations in processing speed, sensory integration, and the ability to sustain performance over extended periods without fatigue or bias. This threshold is increasingly being met due to rapid advances in foundational AI technologies. Large Language Models (LLMs), such as GPT-4 and its successors, now demonstrate sophisticated language comprehension, contextual reasoning, and generative capabilities—enabling them to perform complex tasks such as drafting legal documents, responding to technical inquiries, or managing customer interactions with minimal human intervention. Planning agents, often built upon reinforcement learning or neuro-symbolic architectures, can sequence actions, optimize for long-term objectives, and adapt dynamically to changing constraints. Simultaneously, breakthroughs in multi-modal perception empower AI agents to interpret and integrate data from diverse sources—text, speech, vision, and structured inputs—enabling more situationally aware operation. Complementing these capabilities are memory-augmented models that maintain persistent knowledge over time, supporting continuity across tasks, interactions, and decision cycles. Together, these innovations enable AI systems to operate autonomously in environments where information is vast, response times are critical, and consistency is paramount.

From an economic perspective, the shift toward AI-led operations promises significant efficiency gains and cost reductions. Automating routine, high-volume, and knowledge-intensive tasks allows organizations to scale without linear increases in labor costs, reduce human error, and reallocate skilled professionals to supervisory, creative, or strategic functions. In sectors such as finance, logistics, healthcare, and e-commerce, this transformation supports faster service delivery, improved resource utilization, and the creation of new revenue models driven by real-time, personalized AI services. Moreover, once deployed, AI systems offer non-linear returns on investment, as their marginal operating costs are low relative to the fixed costs of development and integration. Thus, from both a technical and economic standpoint, AI should assume operational leadership in domains where tasks are suited to algorithmic optimization, adaptive learning, and scalable deployment.

One of the most compelling advantages of AI-first systems lies in their ability to automate complex workflows that historically required highly skilled human labor. LLMs can now interpret nuanced instructions and generate domain-specific content, while planning agents can coordinate multi-step processes autonomously. These capabilities allow AI to manage intricate workflows across areas such as legal analysis, medical triage, supply chain optimization, and software engineering—reducing dependency on costly human expertise, minimizing delays, and eliminating inefficiencies. Multi-modal AI further extends this potential by integrating diverse data streams—text, images, audio, and structured data—within a unified automated pipeline, thereby eliminating handoffs between specialized teams. Memory-enabled agents retain institutional knowledge across interactions, reducing the need for repeated onboarding and training. This reconfiguration of traditional workflows drives non-linear scalability and long-term savings while maintaining, or even enhancing, quality and responsiveness.

Strategically, AI-led systems confer a decisive advantage in data-intensive and time-sensitive domains. In finance, AI agents can autonomously analyze market signals, execute trades, detect fraud, and generate real-time risk assessments at a speed and volume beyond human capacity. In logistics, AI can orchestrate global supply chains—dynamically optimizing routes, inventory levels, and forecasts while adapting to disruptions with minimal human involvement. Customer service operations increasingly rely on AI agents capable of resolving high volumes of queries with contextual accuracy and personalized engagement. In intelligence and national security, AI can synthesize multimodal data, identify patterns, and coordinate cyber-defense and surveillance operations across distributed assets. These benefits stem from the ability to orchestrate AI systems that operate continuously, learn from interaction, and respond coherently in complex, dynamic environments. Centralizing operational and cognitive capabilities in AI agents enables organizations to reduce latency, increase adaptability, and outperform competitors—offering a structural advantage that human-led workflows alone cannot match.

While the case for AI-first systems is strong, it is essential to acknowledge the accompanying risks. The misuse of advanced AI—particularly in generating deceptive content, automating cyberattacks, or perpetuating bias at scale—poses real societal challenges. As such, the transition to AI-led systems must be accompanied by strong safeguards, continuous human guidance, and governance mechanisms that ensure alignment with ethical and legal standards.

4 Designing Human-Supported AI Frameworks

Human supervisors play a critical role in the guidance and augmentation of AI systems, particularly in domains where ambiguity, ethical sensitivity, and exception handling are common. Their ability to intervene during unexpected conditions ensures robust system performance beyond the capabilities of pre-programmed responses. Moreover, human ethical supervision is indispensable for evaluating the broader societal implications of automated decisions—especially in high-stakes domains such as healthcare, criminal justice, or autonomous transportation. Supervisors also provide critical support in reasoning about edge cases, where data may be sparse or non-representative, enabling context-aware responses to atypical scenarios. Furthermore, subjective human judgment—grounded in empathy, lived experience, and cultural awareness—remains essential in contexts where algorithmic reasoning alone is insufficient. These human contributions are integral to ensuring the safety, fairness, and accountability of complex AI-first systems.

Governance and auditability are foundational pillars for deploying AI systems responsibly. These pillars ensure transparency, accountability, and alignment with human values throughout the system lifecycle. Transparency enables stakeholders to inspect and understand how decisions are made—requiring not just visibility into system architectures and data pipelines, but also clear documentation of system goals, design assumptions, and limitations. Explainability complements this by making AI outputs interpretable to both technical and non-technical users, thereby fostering trust and facilitating regulatory compliance. Continuous feedback loops support auditability by allowing real-world monitoring, adaptation, and refinement of system performance. These loops enable retrospective evaluations and prospective improvements, forming a dynamic governance structure.

While Human-in-the-Loop (HITL) is traditionally associated with human-led systems, in AI-first frameworks it serves as a supervisory mechanism—where humans intervene selectively to guide autonomous AI agents, rather than directing every decision. HITL principles remain critical in embedding human supervision at key decision points, ensuring that AI systems remain accountable and aligned with ethical, legal, and institutional expectations.

Interface design plays a pivotal role in enabling effective human supervision of AI systems. Rather than focusing solely on execution, interfaces must support supervisory control and situational awareness. User interfaces (UIs) and application programming interfaces (APIs) should be designed to present information clearly, contextually, and hierarchically—enabling supervisors to detect anomalies, understand system states, and make informed decisions. Alerting mechanisms must balance precision and relevance, avoiding both under reporting and user fatigue. Well-calibrated intervention thresholds and structured escalation policies ensure that the right level of attention is directed to the right issues at the right time. These interface design elements foster a collaborative environment where human supervisors act as informed stewards, enabling safe and scalable deployment of autonomous AI systems.

A practical example of this approach is demonstrated by Minerva CQ's Human-in-the-Loop solution for contact center operations [1]. Minerva CQ combines real-time conversational AI with human expertise to enhance customer-agent interactions. The system analyzes live calls, providing agents with timely prompts, relevant knowledge, and decision support. While AI accelerates resolution and augments decisions, humans retain the authority to intervene, override, or refine AI suggestions. This synergy reduces cognitive load, increases customer satisfaction, and enables continuous model improvement through feedback loops—an essential feature in high-volume, high-variability environments like customer support. While Minerva CQ operates as a HITL solution today, it exemplifies how AI-first systems can embed human guidance without compromising autonomy or efficiency.

Table 1 highlights example KPIs reported by contact centers using Minerva CQ's agent-assist solution. These results underscore the practical value of HITL systems in improving both agent performance and customer experience.

Table 1 Contact Center HITL Solution – KPI Improvements

KPI	Improvement
Average Handle Time (AHT)	Reduction of 20–40%
First Call Resolution (FCR)	Increase by 10–20%
Customer Satisfaction (CSAT)	Increase by 10–15%
Agent Assist Accuracy	Improves from 85–90% to 95%+ via feedback
Training Ramp Time	Reduction of 30–50%
Call Deflection	Increase of 10–15% through improved self-service
Agent Turnover Rate	Reduction of 10–25%

These KPI improvements vary by industry and deployment scale but represent consistent gains enabled by effective human-supervised AI systems. Such frameworks exemplify how thoughtful integration of human supervision can enhance performance, maintain accountability, and realize the full potential of AI-first architectures.

5 Risks and Mitigations

The deployment of advanced AI systems within critical decision-making contexts introduces a spectrum of risks that necessitate comprehensive mitigation through systemic, technical, and regulatory interventions. Foremost among these is the potential erosion of effective human supervision, particularly in high-autonomy environments where AI agents may exhibit opaque or unpredictable behavior. To address this, it is essential to implement supervisory mechanisms such as traceable decision logs, formal verification of system behaviors, and interactive simulations that support human understanding and timely intervention. These tools enable human supervisors to examine, test, and approve AI-driven actions either prior to deployment or during real-time operations.

A second major concern involves the amplification of bias and harm—often arising from machine learning models trained on incomplete, imbalanced, or historically biased datasets. Such systems risk entrenching existing social inequities or introducing novel forms of discrimination. Mitigation strategies must therefore include the use of diverse and representative training data, robust fairness auditing tools, and continuous monitoring of system outputs for disparate impacts. Additionally, fallback mechanisms—such as deferring to human judgment or invoking rule-based overrides—should be integrated when predefined fairness thresholds are violated.

The displacement of human labor is another pressing risk, particularly as AI systems increasingly take on tasks that were traditionally performed by skilled cognitive workers. Addressing this challenge requires not only minimizing job displacement but also enabling meaningful workforce transition through targeted reskilling programs.

Organizational restructuring should prioritize human–AI collaboration by redefining human roles around system supervision, design, and strategic guidance, rather than replacement.

Finally, the issue of legal accountability presents a complex challenge, given the distributed responsibility among developers, human supervisors, and autonomous systems themselves. Resolving this requires regulatory innovation, including the development of new liability frameworks capable of apportioning accountability in cases of system failure or harm. This may involve revisiting legal constructs such as product liability, establishing auditable certification schemes for AI systems, and enforcing transparency requirements that ensure decision traceability across increasingly intricate socio-technical infrastructures.

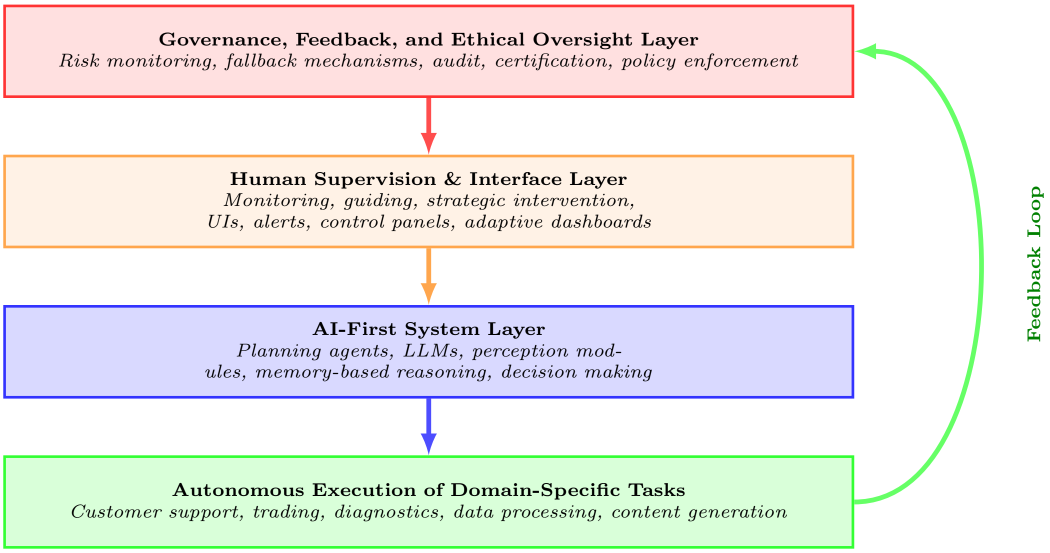
Governance, Feedback, and Ethical Oversight Layer
Risk monitoring, fallback mechanisms, audit, certification, policy enforcement
Human Supervision & Interface Layer
Monitoring, guiding, strategic intervention,
UIs, alerts, control panels, adaptive dashboards
AI-First System Layer
Planning agents, LLMs, perception mod-
ules, memory-based reasoning, decision making
Autonomous Execution of Domain-Specific Tasks
Customer support, trading, diagnostics, data processing, content generation
Feedback Loop

Fig. 1 Layered architecture of AI-first systems with governance and human oversight

6 Case Studies and Prototypes

AI-first approaches are increasingly being deployed in high-impact operational domains, offering both efficiency gains and new challenges in supervision and accountability. Autonomous trading systems, for example, operate at millisecond timescales in global financial markets, using machine learning models to detect patterns, execute trades, and adapt strategies in real time without direct human involvement. While these systems have improved market liquidity and responsiveness, they have also introduced risks such as flash crashes and systemic instabilities, underscoring the need for built-in monitoring and fail-safe mechanisms.

In the domain of self-healing IT infrastructure, AI-driven platforms autonomously detect anomalies, predict failures, and initiate corrective actions—such as re-routing traffic or restarting services—often before human supervisors are even alerted. This has led to significant improvements in uptime and operational resilience but also raises concerns about over-reliance on opaque decision-making in mission-critical environments.

Similarly, AI-led customer service systems now handle large volumes of interactions across sectors such as banking, healthcare, and telecommunications, using natural language processing to interpret queries, resolve issues, and escalate to human agents only when necessary. These systems exemplify human-in-the-loop implementations within AI-first architectures, where AI autonomously manages routine interactions and humans intervene in complex or sensitive cases.

From these examples, several lessons emerge: fully autonomous systems can deliver exceptional performance, but they must be complemented by robust monitoring, explainability, and escalation mechanisms. Moreover, AI-first does not mean AI-only—effective deployment often relies on hybrid architectures in which humans retain meaningful supervisory roles. Looking ahead, responsible AI integration will require adaptive governance frameworks, redefined workforce responsibilities, and continuous alignment of AI systems with evolving ethical, social, and legal norms. The architecture in Figure 1 exemplifies how AI-first systems can operate autonomously while remaining embedded within layers of human supervision and governance, ensuring ethical, adaptable, and task-specialized performance.

7 Strategic Roadmap for Responsible Integration of AI-First Systems

The integration of AI-first systems into complex organizational workflows demands a phased strategic roadmap that aligns technological advancement with responsible governance. A staged approach—structured across short-, medium-, and long-term horizons—enables organizations to manage risk while progressively transforming operational paradigms.

Short Term (1–2 Years)

In the short term (1–2 years), efforts should focus on identifying high-impact, low-risk application areas where AI can deliver immediate value with minimal ethical or safety implications. Typical domains include internal process optimization, IT operations, and customer service triage. During this phase, emphasis should be placed on developing functional prototypes embedded with human guardrails—such as supervision mechanisms, intervention pathways, and interpretable decision-making frameworks. These implementations not only validate technical viability but also serve to build institutional capacity, foster cross-functional collaboration, and cultivate trust among end users.

Medium Term (3–5 Years)

The medium-term phase (3–5 years) emphasizes institutionalizing lessons from early pilots through workforce transformation and governance refinement. This includes designing role-evolution strategies and training programs that empower employees to supervise, audit, and co-work with AI systems. Priorities include the deployment of real-time monitoring tools, feedback loops for continuous learning, and adaptive interfaces that enhance human comprehension and engagement. As AI systems take on more complex and interdependent responsibilities, system transparency and trust become critical enablers of scalable adoption.

Long Term (5–10 Years)

In the long term (5–10 years), the strategic objective shifts toward enabling policy innovation and achieving constrained autonomy in clearly defined domains such as logistics, predictive maintenance, or diagnostic workflows. As seen in earlier case studies—such as autonomous trading systems and self-healing IT infrastructures—long-term integration requires not only advanced capabilities but also robust cross-domain coordination.

To support this, organizations should establish federated supervision frameworks, wherein local teams retain oversight of domain-specific AI agents while centralized governance ensures enterprise-wide alignment with regulatory and ethical norms. This distributed model of human supervision promotes consistency, scalability, and responsiveness across multiple AI-first deployments.

This roadmap outlines a scalable and adaptive strategy for transitioning from pilot deployments to sustainable, AI-first infrastructures that remain aligned with human values and institutional accountability.

Table 2 Strategic roadmap for responsible integration of AI-first systems with impact mapping

Time horizon	**Strategic focus**	**Key actions**	**Impact area**	**Expected outcomes**
Short term (1–2 yrs)	Cautious deployment in low-risk, high-value domains	Build AI-first prototypes with human supervision; ensure interpretability and trust	Process optimization, user trust	Validated use-cases; Increased organizational AI readiness
Medium term (3–5 yrs)	Organizational evolution and trust-building	Redesign roles; scale training; deploy adaptive interfaces; real-time monitoring	Workforce, governance	Higher adoption; Workforce AI fluency; Auditability
Long term (5–10 yrs)	Policy adaptation and constrained autonomy	Federated oversight; coordination of multi-agent AI; regulatory alignment	Cross-domain AI governance	Sustainable AI deployment; Ethical compliance; System-level coordination

8 Conclusion and Next Steps

The transition to AI-first systems represents a profound paradigm shift—one that goes beyond technological innovation to fundamentally redefine the relationship between humans and AI systems. The objective is not to pursue AI supremacy, but to cultivate a symbiotic partnership in which automation augments human insight, and human supervision ensures the ethical and accountable operation of autonomous agents. This vision demands a multi-disciplinary, cross-sectoral commitment to advancing foundational research, accelerating responsible innovation, and constructing adaptive governance architectures resilient to the evolving demands of AI integration.

To shape a future in which AI systems advance collective human interests, design and deployment decisions must be guided not solely by performance metrics but by long-term values such as resilience, fairness, and shared agency. The success of this transition hinges on our ability to embed human values into technical systems from the outset—ensuring future architectures are not only intelligent, but also just, accountable, and societally aligned.

To operationalize this vision, we recommend the following strategic actions:

1. **Establish Cross-Functional AI Governance Councils:** Create interdisciplinary teams comprising technologists, ethicists, legal experts, and domain leaders to guide AI deployment standards, risk management strategies, and policy formation.
2. **Invest in Human-Centered Design and Supervision Mechanisms:** Prioritize systems that enable transparency, interpretability, and human-in-the-loop (HITL) capabilities across varying levels of autonomy.
3. **Scale Workforce Development and Reskilling Programs:** Launch targeted educational initiatives to prepare current and future employees for hybrid roles in AI-augmented environments, focusing on AI supervision, validation, and ethical design.
4. **Promote Open Research and Shared Evaluation Standards:** Support open science and the development of benchmarking frameworks that assess not only technical performance but also societal robustness and ethical alignment.
5. **Develop and Pilot Sector-Specific AI Certification Schemes:** Collaborate with regulatory bodies and industry consortia to prototype audit-ready certification models that build public trust and ensure operational safety and compliance.
6. **Design for Scalable, Federated Supervision Architectures:** Anticipate the complexity of multi-agent systems by developing governance frameworks that enable distributed human supervision and coordinated control across AI-first deployments.

These next steps provide a strategic foundation for realizing AI-first systems that are not only autonomous and efficient but also human-aligned, trustworthy, and adaptable to the evolving needs of society.

Acknowledgements. We would like to express our sincere gratitude to Prof. Franco P. Preparata and Prof. Hal Varian for their support and valuable comments, which greatly enriched the quality of this work.

References

[1] Agrawal, G., Gummuluri, S., Spera, C.: Beyond-rag: Question identification and answer generation in real-time conversations. arXiv preprint arXiv:2410.10136 (2024)

[2] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)

[3] LeCun, Y.: A path towards autonomous machine intelligence. arXiv preprint arXiv:2301.00250 (2023). Meta AI blog

[4] Sutton, R.S.: The Bitter Lesson. https://www.incompleteideas.net/. Accessed: 2025-06-09 (2019)

[5] Marcus, G., Davis, E.: Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon, ??? (2019)

[6] Huang, T., et al.: Ai system design: Technical frameworks for scaling ai-first products. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2) (2021)

[7] Binns, R., Veale, M., Van Kleek, M., Shadbolt, N.: 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018)

[8] Shneiderman, B.: Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction **36**(6), 495–504 (2020)

[9] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., Wallach, H.: Improving fairness in machine learning systems: What do industry practitioners need? In: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019)

[10] Amershi, S., *et al.*: Power to the people: The role of humans in interactive machine learning. AI Magazine **35**(4), 105–120 (2014)

[11] Raji, I.D., Buolamwini, J.: Actionable auditing: Investigating the impact of public audits on algorithmic bias. In: AAAI/ACM Conference on AI, Ethics, and Society (FAT) (2019)

[12] Floridi, L., Cowls, J.: A unified framework of five principles for ai in society. Harvard Data Science Review **1**(1) (2019)

[13] Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, ??? (2014)

[14] Mittelstadt, B.D.: Principles alone cannot guarantee ethical ai. Nature Machine Intelligence **1**, 501–507 (2019)

[15] European Commission: Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Brussels. COM(2021) 206 final (2021)

[16] Suchman, L.A.: Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge University Press, ??? (2007)

[17] Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, ??? (2019)

[18] Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, ??? (2017)

[19] Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, ??? (2021)

[20] Christiano, P.: AI Alignment and Amplification. OpenAI Blog / Alignment Forum. https://www.alignmentforum.org/posts/PaSW5asfAeaPE3xqG/ai-alignment-and-amplification (2018)

[21] Gabriel, I.: Artificial intelligence, values and alignment. Minds and Machines **30**(3), 411–437 (2020)

[22] Zeng, Y., Lu, E., Huangfu, C.: Linking artificial intelligence principles. arXiv preprint arXiv:1812.04814 (2019)

[23] Campbell, S.L., Gear, C.W.: The index of general nonlinear DAES. Numer. Math. **72**(2), 173–196 (1995)

[24] Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. **78**, 74–80 (2000) https://doi.org/10.1007/s001090000086

[25] Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**(2), 321–354 (1995)

[26] Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)

[27] Broy, M.: Software engineering—from auxiliary to key technologies. In: Broy, M., Denert, E. (eds.) Software Pioneers, pp. 10–13. Springer, New York (1992)

[28] Seymour, R.S. (ed.): Conductive Polymers. Plenum, New York (1981)

[29] Smith, S.E.: Neuromuscular blocking drugs in man. In: Zaimis, E. (ed.) Neuromuscular Junction. Handbook of Experimental Pharmacology, vol. 42, pp. 593–660. Springer, Heidelberg (1976)

[30] Chung, S.T., Morris, R.L.: Isolation and characterization of plasmid deoxyribonucleic acid from Streptomyces fradiae. Paper presented at the 3rd international symposium on the genetics of industrial microorganisms, University of Wisconsin, Madison, 4–9 June 1978 (1978)

[31] Hao, Z., AghaKouchak, A., Nakhjiri, N., Farahmand, A.: Global integrated drought monitoring and prediction system (GIDMaPS) data sets. figshare https://doi.org/10.6084/m9.figshare.853801 (2014)

[32] Babichev, S.A., Ries, J., Lvovsky, A.I.: Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon. Preprint at https://arxiv.org/abs/quant-ph/0208066v1 (2002)

[33] Beneke, M., Buchalla, G., Dunietz, I.: Mixing induced CP asymmetries in inclusive B decays. Phys. Lett. **B393**, 132–142 (1997) arXiv:0707.3168 [gr-gc]

[34] Stahl, B.: DeepSIP: Deep Learning of Supernova Ia Parameters, 0.42, Astrophysics Source Code Library (2020), ascl:2006.023

[35] Abbott, T.M.C., *et al.*: Dark Energy Survey Year 1 Results: Constraints on Extended Cosmological Models from Galaxy Clustering and Weak Lensing. Phys. Rev. D **99**(12), 123505 (2019) https://doi.org/10.1103/PhysRevD.99.123505 arXiv:1810.02499 [astro-ph.CO]