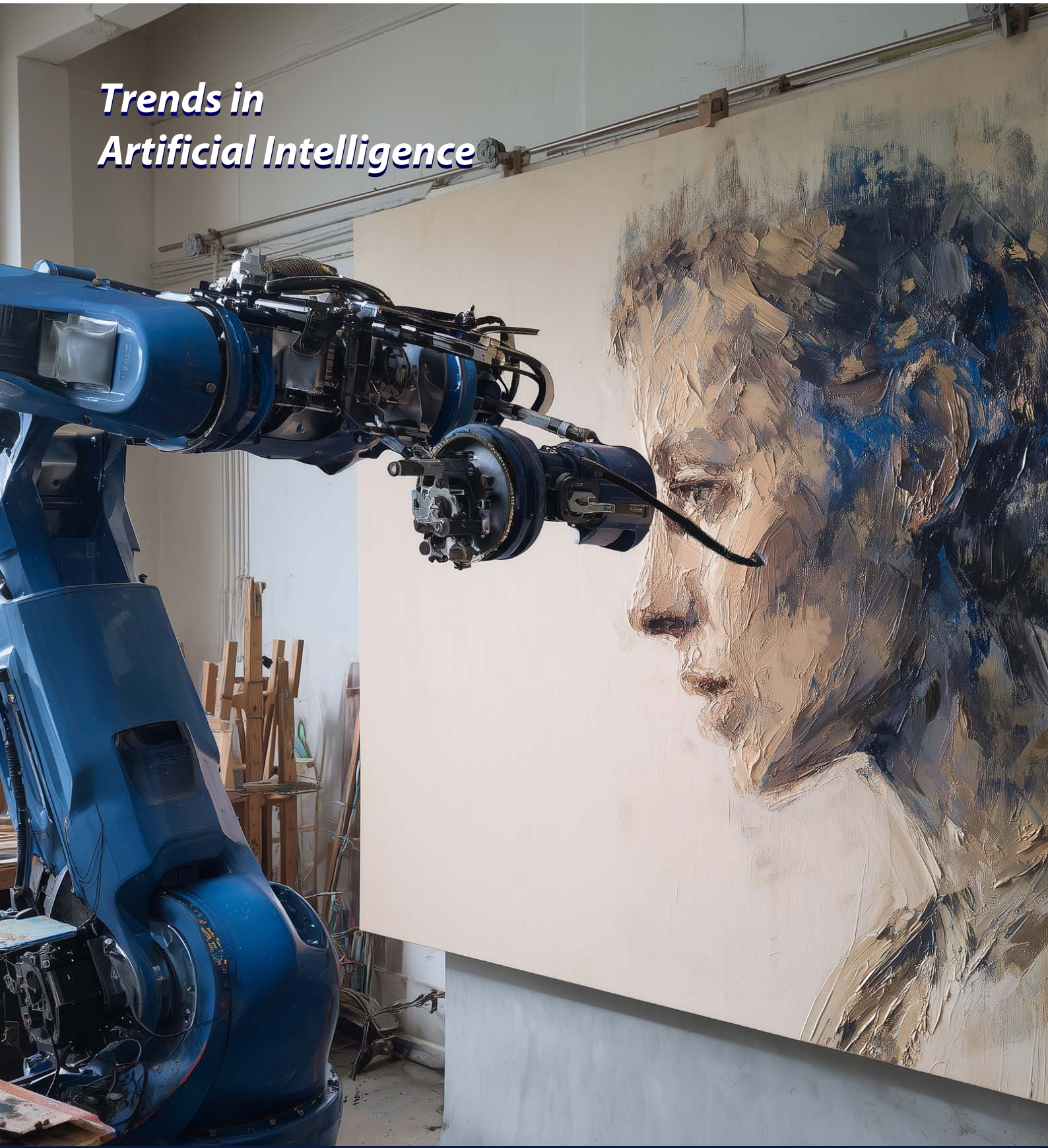
Management Solutions
Making things happen

Trends in Artificial Intelligence



www.managementsolutions.com

Design and Layout
Marketing and Comunicación Department Management Solutions

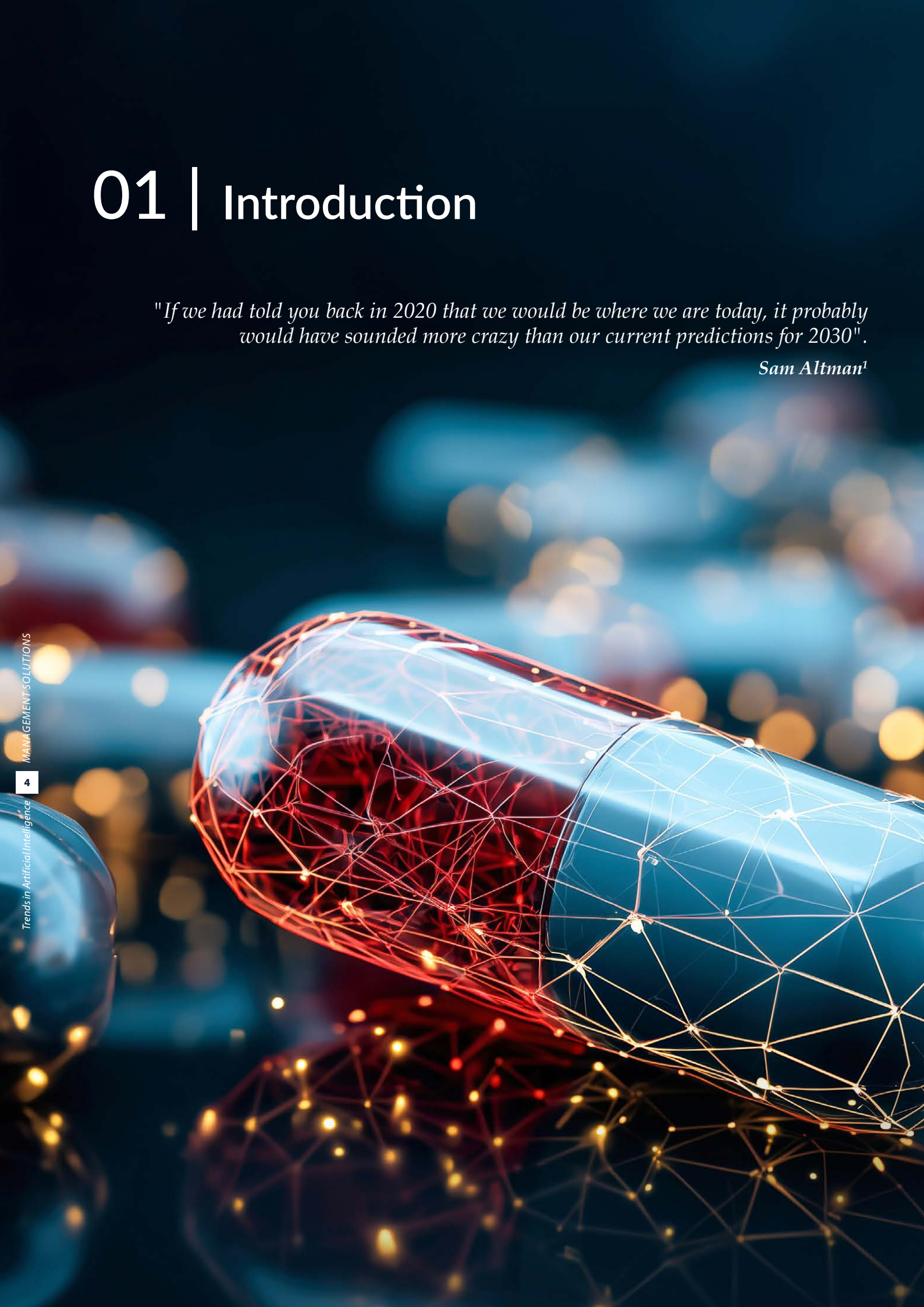
Photographs
Photographic archive of Management Solutions (some images generated with AI)
Adobe Stock

Index

01 | Introduction

"If we had told you back in 2020 that we would be where we are today, it probably would have sounded more crazy than our current predictions for 2030".

***Sam Altman*[1]**



Artificial intelligence (AI) is no longer an emerging technology but a transformative force redefining industries, organizations and societies at unprecedented speed[2]. What seemed like science fiction just five years ago (systems capable of generating text, code, images, music or videos indistinguishable from what humans produce, officially passing the Turing test[3]) is now an operational reality in millions of companies: according to the Stanford AI Index[4], 78 % of organizations were using AI in at least one business function by 2024, up from 55 % the previous year.

1 Samuel Harris Altman (b. 1985), American entrepreneur, founder and CEO of OpenAI.
2 The precise definition of "Artificial Intelligence" is technically ambiguous, philosophically contested and regulatorily significant; the European AI Act provides its own operational definition in Art. 3(1), which nonetheless leaves certain grey areas. For the purposes of this document, the term primarily encompasses generative AI, agentic systems and advanced Machine Learning models.
3 Jones (2025) has shown for the first time in a controlled study that an AI system (GPT-4.5) passes the classical Turing test, being perceived as human in 73 % of conversations.
4 Stanford (2025).

The speed of change doesn't stop: every month new capabilities emerge, every quarter boundaries that seemed far away move, unit costs of AI plummet, and every year it forces us to rethink what we thought we knew about the future of work, competition and business strategy. In the words of Sam Altman[5], CEO of OpenAI.

> *"The cost of using a given level of AI drops approximately 10 times every 12 months [...]. Moore's law changed the world with a 2x improvement every 18 months; this is incomparably stronger."*

And according to Dario Amodei[6], co-founder and CEO of Anthropic:

> *"By 2026 or 2027, we will have AI systems that will be, generally speaking, better than almost all humans at almost everything."*

AI raises strategic questions that go beyond technology and affect strategy, organization, and people:

- How can companies compete when innovation cycles are measured in months?
- How can organizations govern systems that evolve faster than their structures?
- How can they prepare people for jobs that do not yet exist?
- And how can they balance the speed of adoption with effective risk control?

But concrete operational dilemmas also arise:

- Is it better to invest in specific micro-tools that quickly solve bottlenecks, or to go for powerful multi-agent systems that promise consistency and organization-wide impact but require significant investment and carry the risk of rapid obsolescence?
- How can organizations conduct rigorous cost-benefit-risk analysis to prioritize among hundreds or even thousands of pilots?
- And how can they scale prototypes that work in controlled environments but, when deployed at real scale, encounter unexpected costs, emerging hallucinations, and support requirements that overload teams?

The experience of the past few years is beginning to reveal some patterns:

- Effective AI adoption is not just about acquiring technology or launching pilots: it requires organizational transformation, robust governance frameworks, ongoing training, and a deep understanding of technical, regulatory, and reputational risks.
- Organizations that move forward successfully are not necessarily those that invest the most, but those that best integrate technology, people, processes and control.
- The cost of inaction is no longer theoretical: the gap between pioneers and laggards widens exponentially, because every quarter of delay today can translate into years of competitive disadvantage tomorrow.

General productivity tools (securitized enterprise copilots) offer significant improvements immediately, provided they are accompanied by mandatory training and safe use frameworks, as required by European regulations. An OECD review of experimental studies shows substantial average productivity gains from the use of generative AI[7]:

- In writing tasks, average execution time is reduced by 40% and quality increases by 18 %.
- In software development, programmers complete tasks 56% faster.
- In consulting, professionals using AI perform 12% more tasks, complete them 25 % faster and achieve more than a 40% improvement in quality.
- In customer service, professionals supported by AI assistants resolve 14 % more incidents.

The real bottleneck, therefore, is not technical but organizational. Technology is moving faster than internal structures. Friction emerges in slow processes, poorly governed data, overcrowded committees, diffuse responsibilities, and bureaucratic approval cycles. Organizations that do not redesign their internal machinery to enable speed without losing control will be unable to capture the value of AI, no matter how sophisticated the technology they employ. As Gartner puts it:

> *"The enormous potential business value of AI is not going to materialize spontaneously. Success will depend on closely aligned pilots [with the business], proactive infrastructure benchmarking, and coordination between AI and business teams to create tangible business value"[8].*

Two conditions underpin all of this:

- The value that an AI system delivers depends fundamentally on the quality of the data on which it is trained and the quality of the data it operates on in practice: a conversational assistant trained on outdated internal documentation will consistently reproduce those inaccuracies in every response.
- The quality of what a model produces depends largely on the quality of what it is asked to do. The ability to define the problem, provide relevant context and set precise constraints is not a niche technical skill; it is the new operational literacy, and the gap between those who master it and those who do not translates directly into a productivity divide.

Finally, it is critical to manage expectations: AI certainly brings real and measurable value, but it does not immediately replace critical processes or solve structural problems on its own.

5 Altman (2025a).
6 Amodei (2025).
7 OECD (2025).
8 Gartner (2025a).

This paper presents 22 key trends in AI, ranging from capabilities that are already operational to emerging developments that are driving strategic decisions today. It is not intended to be a technical manual or a speculative long-term projection, but rather a rigorous analysis of what is already happening and what is about to happen, designed for decision makers in complex and regulated environments.

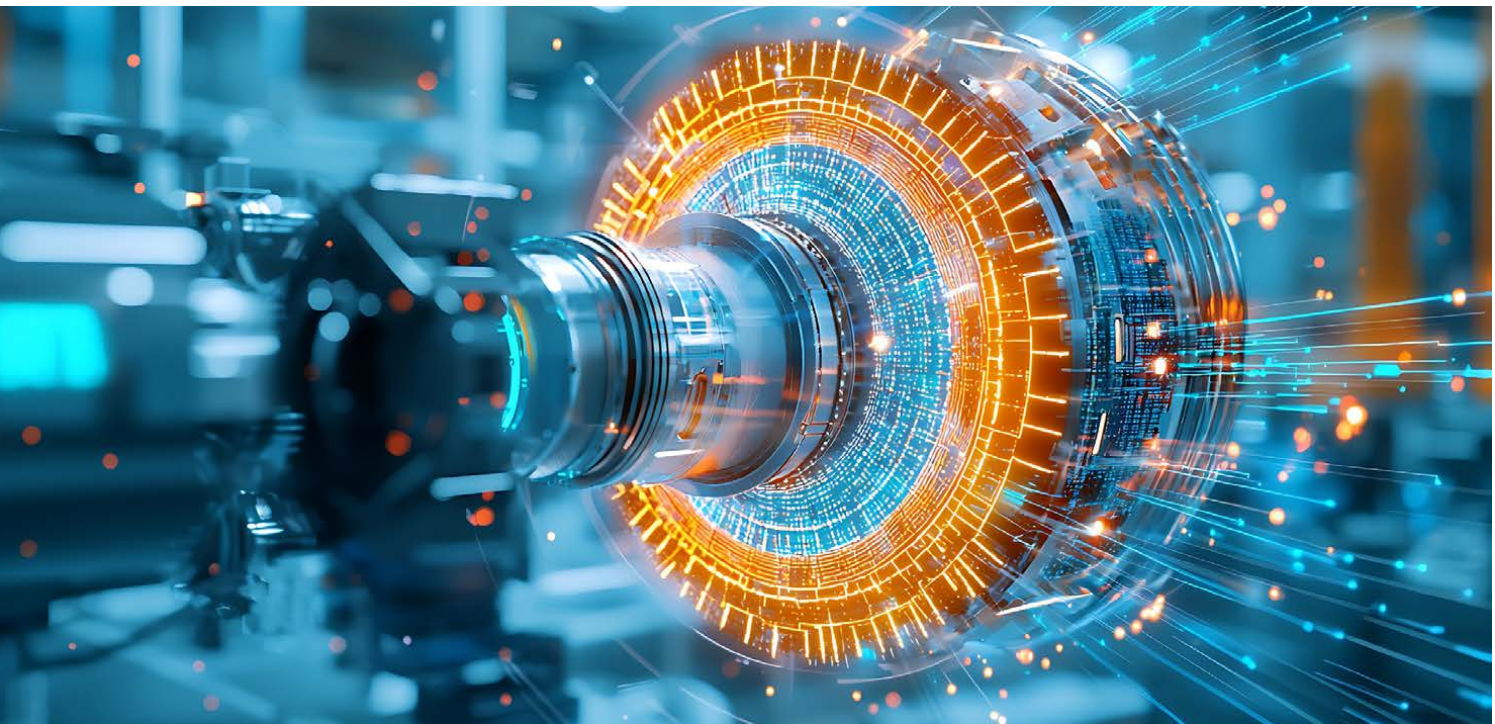
It is structured into four sections:

- **The Technological Explosion of AI** examines capabilities that are already operational and transforming organizations, from the democratization of multimodal generative AI to the rise of agentic systems, including Machine Learning accelerated by generative AI, new approaches to software creation such as vibe coding, and the integration of AI into robotics and physical systems.
- **AI Risks, Regulation, and Safety** addresses the critical challenges of rapid adoption, including technical, legal, and reputational risks; the evolving regulatory frameworks and standards; the emerging conflict between defensive and adversarial AI in cybersecurity; inherent AI vulnerabilities; and ongoing tensions around privacy and intellectual property.
- **AI Governance and Impact on People** focuses on how organizations are responding structurally: creating AI-specific corporate governance models, industrializing deployment through advanced operational practices (MLOps, LLMOps), transforming professional roles and profiles, driving AI adoption across sectors (AI + X), addressing sustainability and social impact, and implementing operational ethical frameworks that go beyond statements of principle; and also addresses the impact of AI on people's daily lives.[9]
- **Frontiers of AI** analyzes emerging developments already shaping strategic decisions: the geopolitics and technological sovereignty of AI, AI-first and AI-only organizations, AI-assisted scientific research, digital twins and simulations of human behavior, Ambient AI and invisible computing, interactions between AI and quantum computing, and artificial general intelligence (AGI) as a strategic horizon that can no longer be ignored.

Finally, a case study is presented: GenMS™ Sybil, a conversational assistant trained using this very paper and developed in a single day. It enables interactive exploration of the trends discussed and illustrates in practice the concepts, architectures, and controls highlighted throughout the document.

This paper is not intended to be exhaustive - AI is evolving too rapidly for that - but rather to provide a solid conceptual framework, concrete examples, verifiable references, and decision-making criteria for navigating an environment of accelerated change with rigor, realism, and responsibility.

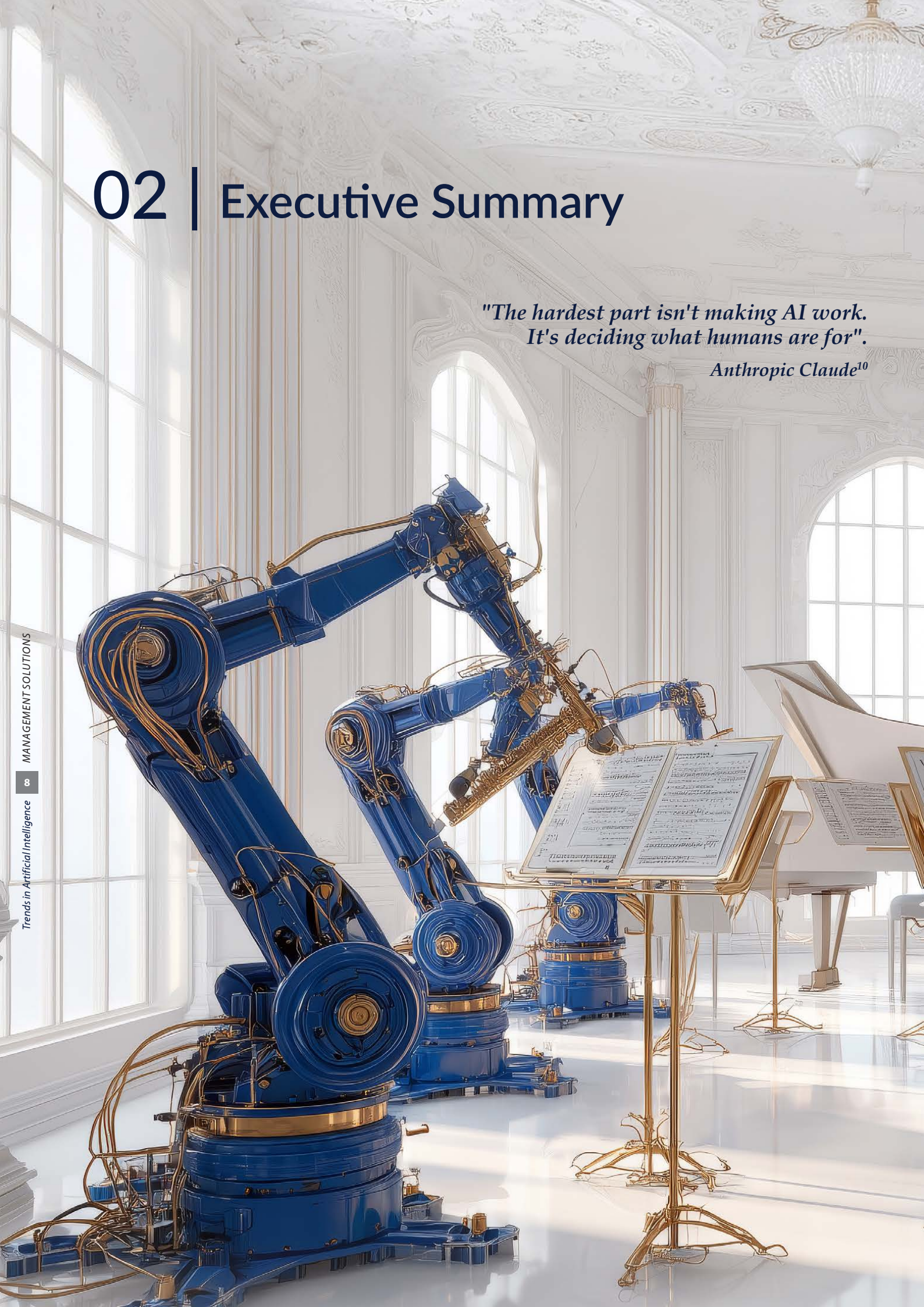
9 iDanae (2Q23), iDanae (1Q24).



02 | Executive Summary

"The hardest part isn't making AI work.
It's deciding what humans are for".

Anthropic Claude[10]



The trends that follow are organised into four areas: the technological explosion of Artificial Intelligence, the risks and regulatory frameworks accompanying its adoption, corporate governance and the impact on people, and the emerging developments already shaping strategic decision-making. For each trend, the most relevant data, documented operational cases and organizational implications are provided. A practical case, GenMS™ Sybil, illustrates the concepts, architectures and controls discussed throughout the document.

10 Claude is a conversational AI system developed by Anthropic (OpenAI spin-off founded in 2021), part of the current generation of large-scale language models (LLMs) along with GPT, Gemini and other frontier systems.

The Technological Explosion of AI

Democratization of multimodal generative AI

Generative AI has become enterprise infrastructure at unprecedented speed: Microsoft Copilot is present in 90% of Fortune 500 companies and ChatGPT is approaching 900 million users. Today's models integrate text, images, audio, video and code into a single conversational architecture, with measurable productivity improvements: scientists publish up to 50% more papers, document processing time is reduced by 80%, and software development speed increases by 56%. This is not incremental optimization: it is a reconfiguration of intellectual work.

This adoption is inevitable. When organizations fail to provide secure corporate tools and adequate training, employees resort to uncontrolled alternatives: up to 35% of the data that professionals upload to unsecured chatbots is confidential. The question is no longer whether to integrate generative AI, but how to do so in a governed way. The European AI Act makes AI literacy a legal obligation from February 2025. Organizations that treat it as a checkbox accumulate risk; those that approach it as a cultural transformation capture a sustainable competitive advantage.

Machine Learning accelerated by generative AI

Classical Machine Learning (ML) remains the backbone of critical applications in multiple industries: credit scoring, fraud detection, demand forecasting, predictive maintenance, etc.

Generative AI does not replace these models; rather, it radically industrializes them by compressing development cycles that previously took months into weeks or even days. Acceleration occurs across all phases: automated generation of predictive variables, creation of technical documentation for regulatory compliance, validation through comprehensive batteries of statistical tests, and automated deployment with continuous monitoring.

A relevant industry development: European banking supervisors already approve ML-based IRB models when institutions adequately justify explainability using techniques such as LIME and SHAP. This dismantles the perception that ML was unfeasible in regulated models. Explainability in ML is not an insurmountable barrier, but it is only partially resolved: current XAI methodologies provide explanations understandable to technical and regulatory audiences, but translating those explanations into terms that a retail customer or an executive committee can readily understand remains an open challenge.

Vibe coding and augmented software creation

Software development has taken a qualitative leap forward: code is no longer written line by line, but in dialogue with systems that interpret requirements, generate complete applications, detect errors and produce tests and documentation automatically. The impact on speed is quantifiable and massive: the task completion rate increases by 26%, projects that used to take months are completed in weeks, and the marginal cost of creating software falls structurally. The democratization is equally profound: business analysts and consultants generate functional prototypes without engineering intermediation.

The flip side is that speed generates hidden risks: invisible vulnerabilities in generated code, model errors replicated at scale, ambiguous specifications that a technician would previously have challenged but which are now executed literally, and a new form of technical debt linked to poorly formulated prompts and implicit architectures. Governing software is no longer governing code: it is governing cognitive systems. This requires versioning prompts repositories, controlling agent autonomy, and tracing which decisions were made by humans versus executed by AI.

Agentic AI and autonomous systems

Agentic AI represents the leap from reactive conversational assistants to autonomous operators that plan, execute complex tasks and act on real corporate infrastructures with full traceability. It already operates in production at massive scale: Deutsche Bank deploys banking agents with an investment of €600 million and savings target of €300 million per year; Ryt Bank processes 80,000 transactions per month with a single conversational interaction; Walmart, Amazon and DHL report productivity improvements of up to 180%.

The real challenge is not building agents but governing and scaling them. Technical scalability requires interoperability standards such as MCP (Model Context Protocol), which eliminate the technical debt of proprietary integrations and turns each tool into an asset reusable by any agent. Organizational scalability requires effective human oversight, explicit limits on what each agent can execute, and rigorous cost control: viable prototypes become economically unsustainable systems without these safeguards designed from the outset. To this we must add a structural constraint: human supervisory capacity has a ceiling, and once exceeded, supervision becomes nominal—more dangerous than its absence, given the false sense of control it creates.

AI in robotics and physical systems

Industrial robotics has crossed a qualitative threshold: today's robots perceive their environment in real time, interpret instructions in natural language, adapt to changes without reprogramming, and learn from every interaction. Humanoid robotics has made the ultimate leap from the lab to the factory floor: Figure AI completed an eleven-month deployment at BMW in 2025 where two robots worked 1,250 hours and contributed to the production of 30,000 vehicles; Tesla plans to manufacture one million Optimus units annually in 2026 at less than $20,000 per unit; Boston Dynamics operates its electric Atlas via Large Behavior Models with industrial pilots underway.

The advantage is structural: robots can operate 24/7 without fatigue and with predictable recurring costs. The risks are equally structural: concentrated impact on repetitive manual jobs, dependence on proprietary ecosystems, accelerated technological obsolescence, and the need for robust safety frameworks with effective human oversight, even in nominally autonomous operations.
Beyond manufacturing, humanoid robotics is opening a second front: the care of older and dependent individuals, bringing its own strategic, ethical and regulatory implications.

AI Risks, Regulation and Safety

Risks of AI

AI does not introduce substantially new risks: it amplifies them. An algorithmic bias is a human bias systematized and replicated millions of times; an information leak from misuse of a chatbot is, in the end, an information leak. The difference is in the speed of propagation, the scale of the impact and the difficulty of containment.

The key phenomenon is non-linear amplification: a minor glitch (a poorly designed prompt, a misconfiguration of permissions) can escalate in minutes and simultaneously affect processes, customers, regulators and reputation. A customer service model that leaks confidential information in 0.01% of conversations generates 10 incidents per day in a system of 100,000 interactions, each with regulatory, contractual and reputational implications, before the pattern is detected.

Risks materialize in four dimensions: (1) security and compliance (e.g., prompt injection, data leaks); (2) quality and reliability (e.g., hallucinations, explainability, model drift, vendor lock-in, cost escalation in agentic systems); (3) ethics and automated decisions (e.g., amplified biases, accountability gaps in distributed causal chains); and (4) social impact (e.g., erosion of critical capabilities, employment transformation, environmental footprint).

Two specific economic risks also emerge: although the unit costs of Artificial Intelligence are structurally declining, poorly governed agentic systems can trigger total costs in a non-linear way, and there is uncertainty as to whether massive investment in AI will have the expected return: Gartner forecasts that more than 40% of agentic projects will be cancelled before 2027 for these two reasons.

AI regulation, oversight and standards

Unlike previous technology cycles, AI is being regulated in parallel to its mass deployment. Europe leads with the AI Act (EU Regulation 2024/1689), the first comprehensive legal framework on AI: it classifies systems by risk level, imposes structural obligations on high-risk ones (documented risk management, traceability, human supervision, prior compliance assessment) and sets penalties of up to €35 million or 7% of global turnover, surpassing GDPR. The supervisory architecture (AI Office, national authorities, and the AI Board) is currently being established, with Spain a pioneer in designating a national authority (AESIA).

The rest of the world shows no convergence. The United States maintains a fragmented sectoral approach with no federal equivalent to the AI Act, and focuses on global supremacy in AI; China integrates AI into a strategy of digital sovereignty with compulsory licensing and data control; the United Kingdom is committed to pro-innovation principles without horizontal legislation; Brazil is advancing a model similar to the European one pending parliamentary approval.

In parallel, technical standards such as ISO/IEC 42001 or NIST AI RMF are forming the operational basis of compliance programs. For global organizations, this fragmentation translates into multi-level AI architectures designed to simultaneously reconcile divergent requirements across jurisdictions.

AI and cybersecurity

Cybersecurity has become a battle of AI versus AI. More than 28 million AI-powered cyberattacks were recorded in 2025, a 47% year-over-year increase, and 87% of organizations experienced at least one. The vectors are qualitatively new: hyper-personalized phishing generated by LLMs with success rates of 54% versus 12% for traditional phishing; polymorphic malware that rewrites its own code every 15 seconds to evade signature detection; audio and video deepfakes that impersonate executives in BEC attacks; and dark LLMs such as WormGPT or FraudGPT marketed on the Dark Web, with technical support included.

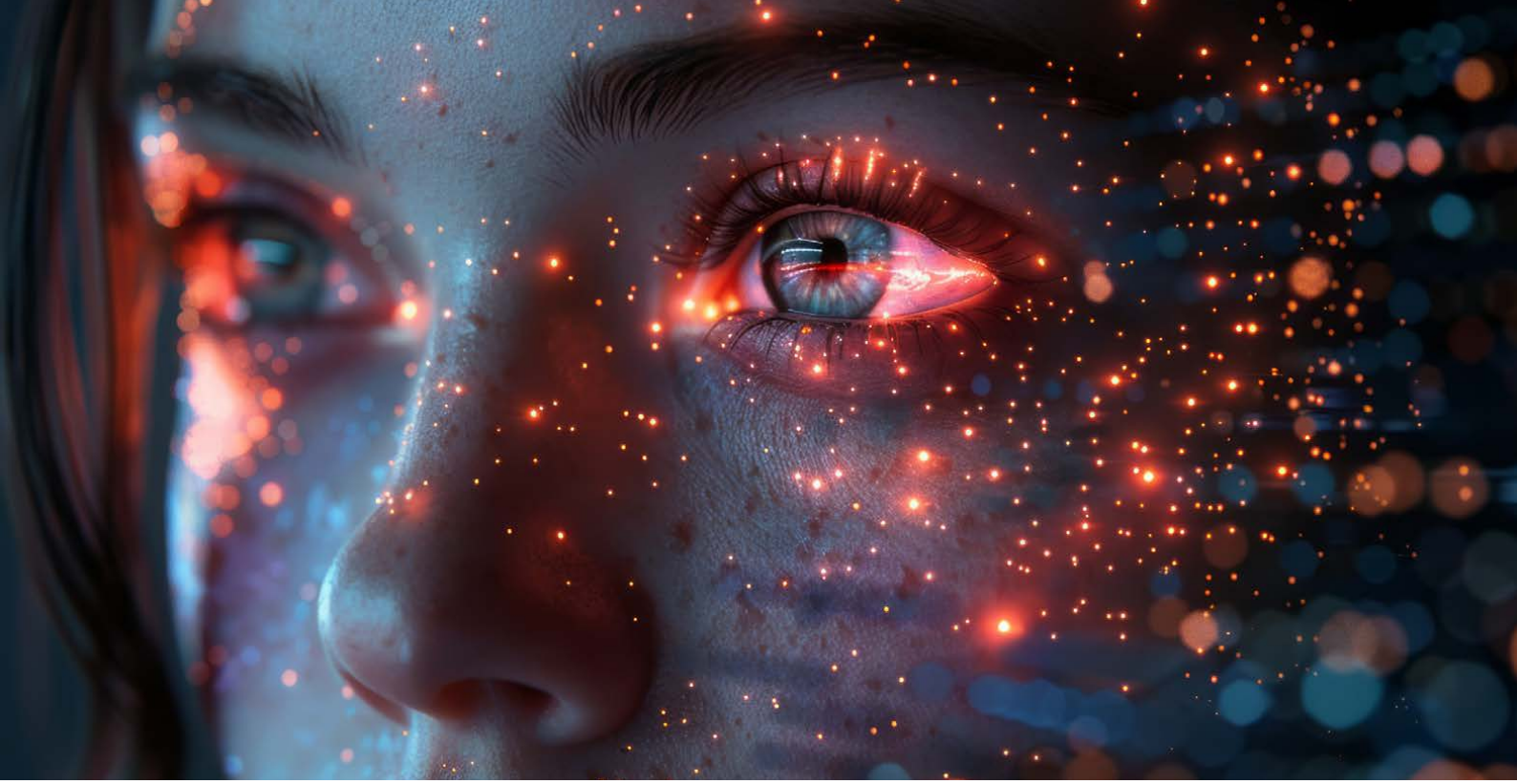
The defensive response is equally sophisticated: UEBA systems analyzing billions of daily events achieve detection rates of 98%, AI-enabled SIEM/XDR/SOAR platforms reduce false positives by up to 95% and shorten containment cycles by 80 days, and organizations deploying defensive AI reduce the average cost of breaches by $1.9 million. But a structural asymmetry remains: the advantage no longer stems from simply having AI, but from the sophistication of the models and the speed with which threat intelligence is updated.

A third dimension emerges that traditional frameworks do not consider: AI systems themselves are attack surfaces, vulnerable to data poisoning, adversarial evasion, and prompt injection, creating a meta-layer of risk that requires its own controls.

AI, privacy and intellectual property

The operational logic of LLMs fundamentally clashes with privacy and intellectual property frameworks. In terms of privacy, each stage of the LLM lifecycle carries distinct risks: the unintentional storage of personal data that can be extracted via prompts; the potential re-identification of individuals from seemingly anonymized outputs; and feedback loops where user interactions with chatbots are incorporated into model retraining without consent. This creates a structural incompatibility with GDPR: LLMs require vast amounts of data, violating the principle of data minimization; they cannot be selectively de-trained, conflicting with the right to be forgotten; and their architectures are opaque, undermining transparency requirements. The EDPB therefore concludes that Data Protection Impact Assessments (DPIAs) are mandatory in most cases. While technical mitigations such as differential privacy, federated learning, and retrieval-augmented generation (RAG) exist, they come with trade-offs in accuracy, computational cost, or functionality.

In intellectual property, the core debate over whether training models on protected content constitutes "fair use" or constitutes massive infringement remains unresolved in court. Over 72 lawsuits are currently active against AI companies, including The New York Times vs. OpenAI, Getty Images vs. Stability AI, and record labels vs. Anthropic. Ownership of AI-generated outputs is similarly unclear: if human intervention is insufficient, the content falls into the public domain, yet the threshold of what counts as "sufficient" is undefined. Underpinning all of this, WIPO warns that the global rights management infrastructure, built for human-scale creation, strains under the trillions of outputs AI generates every day.



AI Governance and Impact on People

Corporate governance of AI

AI overwhelms traditional governance frameworks: it makes decisions without human intervention, produces non-deterministic outputs, operates through opaque internal processes, and relies on external providers whose models evolve without direct organizational control. Governance structures designed for predictable technologies are too slow, lack the necessary expertise and are not equipped to manage this uncertainty.

The emerging organizational response follows a hub-and-spokes model, combining a central Center of Excellence with decentralized teams embedded in lines of business, alongside an AI Risk or AI Governance coordination function that orchestrates assessments across specialized units. Currently, 26% of large organizations have a CAIO, CDAIO, or equivalent role; at smaller scales, positions such as AI Risk Manager or AI Ethics Officer are appearing, though without standardization.

Real governance does not happen in the AI Committee itself, but in the AI Working Group that prepares it: this is where positions are negotiated, tensions between speed and control are resolved, and the agreements that the committee will formally approve are built. Regarding risk frameworks, organizations typically do not start from scratch; instead, they enhance existing frameworks by adding AI-specific chapters to areas such as Model Risk, Supplier Risk, Data Protection, and Compliance. Similarly, while the regulatory classification under the AI Act is necessary, it is not sufficient; organizations complement it with more detailed internal taxonomies that account for reputational impact, process criticality, supplier maturity, and other factors.

Industrialization of AI (MLOps, LLMOps)

The main bottleneck in AI adoption is not algorithmic, but operational: promising pilots in experimental environments often fail to reach production, or when they do, they suffer performance degradation, generate unexpected costs, and introduce unmanageable risks.

MLOps addresses this challenge by providing standardized processes for building, deploying, and operationalizing models reliably throughout their lifecycle. LLMOps extends these practices to generative models, managing their unique characteristics - nondeterministic behavior, prompt-related risk surfaces, hallucinations, and costs that can scale unpredictably.

Industrializing AI means creating the operational infrastructure that makes models reliable, auditable, and sustainable in real-world production. This includes continuous validation with human oversight, real-time monitoring of costs and behavior, controlled deployment pipelines, and full traceability as required by the AI Act. Without this operational layer - provided by MLOps and LLMOps - governance frameworks risk remaining mere statements of intent.

Upskilling, reskilling and new professional roles

A key challenge for organizations is having the right capabilities to design, deploy, operate, and govern AI systems. AI talent can be grouped into three categories: technical profiles (ML engineers, data architects, LLMOps specialists, etc.), hybrid profiles that bridge technical expertise with business needs, and governance and control profiles (AI Risk Manager, AI Ethics Officer, AI Compliance Lead, etc.)

An empirical analysis of 16 large European and US organizations shows a clear convergence around this core set of roles. The main variation lies in which organizations have formally institutionalized the most specialized profiles versus those that maintain them informally, leading to gaps in control and scalability.

The talent market shows a structural imbalance: demand systematically exceeds supply across nearly all profiles, with the partial exception of Data Scientists. The shortage is most acute in production roles (MLOps, LLMOps) and governance/control roles, where the combination of technical complexity, required seniority, and rising regulatory requirements outpaces the market's capacity. Outsourcing alone cannot close this gap; internal upskilling and reskilling are therefore the inevitable structural levers for organizations.

AI and sector transformation (AI + X)

AI is no longer a technology adopted on a sector-by-sector basis; it has become a cross-cutting layer of intelligence, integrated simultaneously across all domains of activity, nevertheless some of the most significant advances continue to be driven by specific sectors each with its own underlying dynamics. The IMF estimates that 40% of global employment is exposed to AI, with figures exceeding 60% in advanced economies. The ILO notes that, for now, the impact is concentrated on specific tasks rather than entire occupations, implying job reconfiguration rather than wholesale substitution. The OECD classifies sectors by their "AI intensity" and observes that even the least-digitized sectors are increasing their exposure, with cross-domain acceleration effects.

Operational applications already span all sectors: AI systems achieving diagnostic accuracy comparable to specialists in radiology and dermatology; adaptive tutors delivering personalized learning at scale; predictive maintenance and advanced robotics in industry; fraud detection and document automation in finance; and text, image, and music generation in creative industries. What matters is not the individual applications, but the pattern: competitive advantage no longer comes from applying AI to isolated functions, but from integrating it as a cognitive infrastructure across the entire value chain.

AI in personal and everyday life

Generative AI has reversed the historical paradigm of technology adoption. Unlike cloud, ERP, or CRM systems—which originated in corporate environments and later spread to consumers—AI first entered personal life. In the EU, 25.1% of the population uses it for personal purposes, compared to only 15.1% in work contexts. Among students over 16, 75% use AI regularly, while only 12.5% of retirees do. The resulting generation gap of 53.6 percentage points far exceeds differences by education or income. Organizations are not driving this transformation; instead, they are reacting to capabilities employees already possess and use unofficially, creating "shadow AI" exposures that most companies are not yet controlling.

Mass adoption coexists with deep ambivalence. Globally, 66% of people expect AI to have a significant impact on their daily lives in the coming years, yet 51% of U.S. adults report feeling more concerned than excited. Acceptance varies widely - by as much as 110 percentage points -depending on the use case. Both the general public and experts share a common frustration: 55% want more control over how AI affects their lives, but fewer than 25% feel they have it. Access asymmetry adds another dimension: those who integrate AI as an everyday cognitive tool gain advantages in learning, productivity, and creativity at a pace that disconnected groups cannot match.

AI, sustainability and social impact

The relationship between AI and sustainability is bidirectional and tense. On the one hand, AI acts as a transition accelerator: it optimizes electricity grids, improves renewables integration, refines climate modeling, and can reduce emissions on the order of 1,400 Mt CO_2eq annually by 2035 in wide adoption scenarios.

On the other, its own infrastructural footprint is growing and difficult to ignore[11]: data centers will consume 945 TWh per year by 2030 (equivalent to Japan's electricity consumption today), training of frontier models grows more than 2x per year in required power, and the largest individual runs could demand between 4 and 16 GW by 2030, on the same magnitude as several nuclear power plants. CO_2 emissions associated with data centers could reach 300-320 Mt per year by 2030 if additional electricity continues to rely on fossil fuels.

The distributional dimension adds another layer of complexity. Economies with higher technology density capture the efficiency benefits earlier, while others bear the transition costs without accessing the gains. The geographic concentration of computational capacity also reconfigures strategic dependencies and access to technology on a geopolitical scale. Evaluating AI in terms of sustainability therefore requires explicit metrics of energy and water consumption, transparency about the location of deployment, and analysis of the distribution of impacts, not just their aggregate magnitude.

AI ethics and philosophy

Since 2017, more than 245 AI ethics frameworks have been issued, yet the sheer proliferation of principles has not resolved - or even significantly reduced - ethical challenges. The real operational risk lies in the gap between stated principles and the difficulty of monitoring actual AI behavior. Closing this gap requires a shift from declarative ethics to operational ethics.

Working AI ethics frameworks share six core components: (1) a governance structure with clearly defined responsibilities; (2) individualized impact assessments for each system, proportional to its autonomy and potential consequences; (3) continuous bias management rather than one-off audits; (4) differentiated explainability tailored to the audience - regulators, customers, or affected employees; (5) accessible escalation and whistleblowing channels; and (6) periodic review of the framework as models evolve.

In 2026, Anthropic published its Constitution, the first document from a frontier AI laboratory that encodes principles and values directly into model training, aiming for the system to internalize the reasoning behind each principle, not merely follow rules.

Underlying this effort is a question that current regulatory frameworks are not designed to address: what kind of entity are we governing? A credit scoring system, a conversational assistant, and an autonomous agent negotiating contracts may fall under the same regulatory risk category, yet each carries fundamentally different ethical obligations. Anthropic has publicly acknowledged that Claude "may possess some form of consciousness," becoming the first frontier lab to admit it cannot answer with certainty what it has created. This raises profound ethical and philosophical questions for which no answers currently exist.

11 And these projections already incorporate the continuous gainsin hardware energy efficiency.

Frontiers of AI

Geopolitics and technological sovereignty of AI

AI has become strategic state infrastructure. Sovereignty now operates across several layers: hardware (ASML is the world's sole supplier of EUV lithography, without which the manufacture of advanced chips is not possible; TSMC manufactures more than 90% of those chips, while NVIDIA controls over 85% of the GPU market used for training), infrastructure (Amazon Web Services, Microsoft Azure, and Google Cloud Platform together account for roughly two-thirds of global computing capacity), and talent, whose mobility effectively turns migration policy into technology policy. The strategic question is not how many layers are controlled, but which ones are critical to one's mission.

Three models compete with different logics: The United States combines private primacy with the greatest technological export controls since the Cold War; China, which has demonstrated with DeepSeek that hardware containment has limits, pursues declared self-sufficiency across the entire value chain by 2030; Europe exerts influence through regulation - the "Brussels effect" forces global products to adapt to its standards - but maintains deep infrastructural dependencies. The result is partially incompatible technoblocks where full decoupling would force third parties to choose sides at prohibitive costs.

For organizations, the implication is straightforward: dependence on a single foundational model provider is already a strategic risk, not just an operational one. Multi-model and multi-cloud strategies are today the corporate equivalent of diversifying sovereign dependencies.

AI-first and AI-only organizations

Three stages define the spectrum. AI-enhanced organizations (current majority) use AI to improve existing processes. AI-first organizations design their processes from AI capabilities: Midjourney and Cursor exceed $500 million in revenue with less than 163 and 50 employees respectively - ratios of more than $3 million per employee that exceed historical industry benchmarks by an order of magnitude; MYbank approves credit to 50 million SMEs without human intervention in less than a second.

AI-only organizations (with no humans in core operations) do not yet exist: in regulated sectors, they are prevented by regulations; in less regulated sectors, they are constrained by the error rates of agents in extended workflows and by the absence of clear legal liability mechanisms. The strategic question is not whether they will exist, but who will build them. They will probably not evolve from existing organizations, but as new entities without operational heritage. This pattern can already be seen in examples such as Ping An - which developed eleven independent startup subsidiaries, five of them publicly listed - and DBS Bank with its digital bank Digibank.

Digital twins and the simulation of human behavior

Digital twins originated in aerospace engineering as tools for modeling deterministic physical systems such as turbines, airframes, or electrical networks. Their historical limitation was epistemological rather than technological: complex systems - cities, markets, organizations - cannot be modeled in the same way because their behavior emerges from interactions among agents rather than being deduced from their components. More data and greater computational power do not resolve this problem.

Large language models have introduced a discontinuity at this frontier. In 2023, a team at Stanford University created 25 agents with identities, memory, and social relationships built on LLMs; their collective behaviors emerged without being explicitly programmed. In 2024, the same researchers replicated the responses of 1,052 real individuals in standardized surveys with 85% accuracy -comparable to the variability of the individuals themselves. The startup Simile, which raised $100 million in February 2026, is already commercializing digital twins of individuals to simulate customer behavior. As a result, the $142-billion global market research industry faces potential structural disruption.

The next step is to simulate not thousands of individuals but entire populations in real time, allowing policymakers or organizations to anticipate how a society might respond to a tax reform or regulatory intervention before implementing it. Such a capability would have no historical precedent - and no existing governance framework to regulate it.

Ambient AI and invisible computing

Ambient AI operates without being invoked: it continuously observes context, infers needs, and acts proactively; the interface disappears. This has become possible thanks to the simultaneous maturity of three elements: small models capable of running locally on devices without reliance on the cloud; dense networks of physical and biometric sensors; and LLMs capable of reasoning about heterogeneous context in real time.

One of the best-documented applications is the use of clinical ambient scribes: systems that listen to doctor-patient conversations and automatically generate clinical documentation. A randomized trial at UCLA evaluated two such platforms across 238 physicians and more than 72,000 patient encounters, finding measurable reductions in documentation burden and burnout. Yet this remains a relatively bounded application. What's coming - workspaces that infer occupants' attentional states, wearables that alert users before symptoms become consciously perceived, and agents that manage schedules and resources within defined parameters - will make current cases seem rudimentary.

These developments create structural tensions. Privacy faces a new challenge: the appetite for biometric and behavioral data in these systems renders conventional informed consent inadequate. Errors become invisible: in an invoked system there is a request against which the response can be compared; in an ambient one, there is not. And the AI Act, designed for systems with an intended purpose, doesn't address AI that continuously observes and adapts.

Interaction between AI and quantum computing

AI and quantum computing are distinct technologies that intersect at three points. The first two are medium-term prospects: quantum computing could speed up the training of AI models (which is essentially an optimization problem over extremely large parameter spaces) and run certain ML algorithms more efficiently, particularly for sorting and combinatorial optimization problems. Current evidence does not justify the hype - in many cases, classical systems with good data remain competitive - and the necessary hardware will not be available at commercial scale before the end of this

decade, and likely beyond, as Artificial Intelligence models are scaling faster than progress in quantum hardware.

The third crossover point is different: it is not a future opportunity but a present threat to the infrastructure on which all AI deployed today operates. Virtually all the cryptography that protects digital communications: banking transactions, medical records, regulatory communications, and channels between AI systems, is based on mathematical problems that a sufficiently powerful quantum computer could solve with ease. State actors are already capturing encrypted data today to decrypt it when that capability arrives, a strategy known as "harvest now, decrypt later." The National Institute of Standards and Technology (NIST) published the first quantum-resistant cryptography standards in 2024. Organizations with sensitive, long-lived data should start migration now: in complex organizations the process takes years, and waiting for the relevant quantum computer to exist would mean starting too late.

Artificial General Intelligence (AGI) as a strategic horizon

AGI designates AI capable of performing the full range of cognitive tasks humans can perform, with the ability to generalize across domains. There is no consensus about whether it already exists: in February 2026, the journal Nature published two papers by leading researchers that reached opposite conclusions. This highlights a key point: "general intelligence" is a continuous concept without clear thresholds. The strategically relevant question is therefore not philosophical but functional: when can a system autonomously complete entire cycles of high-value cognitive work? In several domains, that threshold has already been crossed.

What comes next follows a logic of cumulative escalation: from tool to agent, from agent to environmental infrastructure. In parallel, AI is improving itself in a self-reinforcing loop, a process that is driving progress toward over-exponential growth. The structural consequence is unprecedented: the upper limit of reasoning available on the planet, which since the first hominids has been human intelligence, is being displaced.

The determining variable is not access to the best models, which will increasingly become commoditized, but the speed of organizational absorption: redesigning processes, transforming roles, building effective governance. The largest gains appear not where AI simply replaces tasks, but where it reorganizes entire processes. At the same time, the greatest systemic risk is that cognitive capacity is concentrated in a few actors, whose advantage is self-reinforced by the very feedback loop that speeds up overall progress. Institutional responses today lag far behind the pace of this transformation.

To treat AGI as a strategic priority, it is not necessary to settle the philosophical question of what it is or whether it already exists; what matters is recognizing that its consequences are already unfolding today.

Case study: GenMS™ Sybil

GenMS™ Sybil was specified, built, secured, validated and deployed in a single day, fully following the LLMOps cycle. It is a public conversational assistant based exclusively on this document, designed from the outset under regulatory compliance, privacy and security criteria, which conditioned the architecture from the specification phase.

The process covered all phases: deliberate delimitation of the corpus to avoid intellectual property risks; complete technical specification (architecture, operational limits, quality and security metrics) developed through structured interaction with an LLM; continuous validation including human review, stress testing and red-teaming, complemented by GenMS™ Atlas on dimensions such as bias, robustness, privacy and compliance; code generation and auditing within the same cycle; and deployment with active monitoring of costs, traceability and usage control.

Architecture decisions were explicit: full context versus RAG to preserve global consistency; prompting instead of fine-tuning to ensure traceability; proprietary boundary model to maximize stability; independent sessions to meet the minimization principle. The multi-page system prompt encodes the actual system guardrails.

This case does not describe trends: it executes them. It demonstrates that the industrialization of generative systems is feasible when an organization has method, technical expertise, and built-in governance.

03 | The Technological Explosion of AI

"It's going to be 10 times bigger than the Industrial Revolution - and maybe 10 times faster".

Demis Hassabis[12]

AI systems have crossed critical thresholds in recent years: they no longer just communicate, they execute; they no longer just suggest, we start delegating decisions to them; and they no longer just operate in digital environments, they act on the physical world. What began as a personal productivity tool has evolved into operational infrastructure capable of automating complex, end-to-end cognitive processes. This section examines five technological trends that are redefining the limits of what is possible. They are not describing the future: they are describing the operational reality of organizations already capturing structural competitive advantages through AI.

12 Demis Hassabis (b. 1976) is CEO of Google DeepMind and 2024 Nobel laureate in Chemistry.

Democratization of Multimodal Generative AI

The transformation of intellectual work

In less than five years, generative AI has gone from technology experiment to business productivity infrastructure. Tools such as Microsoft Copilot, ChatGPT Enterprise and Claude Enterprise have been massively deployed in workplaces around the world: according to Satya Nadella[13], Microsoft Copilot is present in 90% of Fortune 500 companies and has surpassed 150 million users[14]; and ChatGPT is estimated[15] to be approaching 900 million users.

This speed of adoption is unprecedented in enterprise technologies: neither the cloud, nor mobile devices, nor collaborative platforms reached this penetration with such speed.

What is happening is not an incremental improvement of existing tools, but a reconfiguration of how intellectual work is done. Tasks that used to take hours (writing reports, analyzing lengthy documents, generating code, creating presentations, synthesizing scattered information) can now be completed in minutes[16], producing an initial useful output through conversational interaction with AI systems.

Effective adoption cases

Productivity improvements are already visible. To cite a few cases:

- In standard office tasks, productivity gains of 10% to 13% in document editing, 11% reductions in email processing time, and 23% faster resolution of IT security incidents have been measured[17] .
- Studies show[18] that scientists using large language models (LLMs) publish up to 50% more scientific papers than before using these tools[19] .
- Empirical studies show[20] that business process automation with generative AI can reduce corporate document processing time by more than 80%, while also lowering error rates.
- In customer service[21], the introduction of conversational assistants based on generative AI enabled 14% more incidents to be resolved compared to traditional processes.
- A systematic review[22] of AI adoption in the workplace finds consistent increases in efficiency and productivity across a range of tasks following the adoption of generative AI, while also warning of certain risks associated with its use.

13 Nadella (2025).
14 Although the market for enterprise assistants continues to evolve at pace, with new competitors steadily consolidating their position.
15 FirstPageSage (2025).
16 The time required to reach the final outcome logically depends on the refinement iterations, the complexity of the task and the rigour of the verification process—factors that the professional cannot and should not fully delegate to the system.
17 Stanford (2025).
18 Kusumegi (2025).
19 iDanae (2Q23).
20 Jeong (2025).
21 OECD (2025).
22 Yuan (2025).

Multimodality as a qualitative leap forward

Current models integrate text, images, audio, video and code into a single general-purpose architecture, a trend that coexists with the parallel development of highly specialized models that outperform general-purpose systems in specific domains. A user can upload a picture of a financial scorecard drawn on a whiteboard and receive the full code needed to reproduce it; can dictate a complete presentation while the system simultaneously generates slides with graphs and charts; can provide a regulation and obtain a podcast explaining it, or ask the system to analyze a video of a meeting and extract decisions, commitments, and deadlines, for example.

This multimodal convergence is not minor: it redefines the notion of what tasks are automatable. Tasks that previously required multiple specialized tools (e.g., audio transcription → text analysis → graphics generation → report writing) are now solved in a single conversational interaction. The capabilities of these systems continue to increase quarter by quarter with no signs of slowing (Fig. 1), implying that the impact on speed and accessibility will continue to amplify.

Democratization: from technical experts to non-technical profiles

The conversational interface eliminates barriers to entry and goes beyond the concept of "citizen data scientist" (non-technical professionals capable of performing basic analysis with visual tools) to deliver analysis, code generation and advanced processing capabilities directly to end users, without the need for technical training.

This democratization, however, has a double edge: on the one hand, it frees productive capabilities previously limited to specialists; on the other, it spreads risk: now any employee can, without technical supervision, generate content, code or analysis that the organization could use in critical decisions. The key question is not whether to democratize access, but how to govern use on a massive scale.

The problem of verification at scale

Generative AI produces, by design, outputs that are plausible but not necessarily correct, and the root of this issue is structural: these systems are statistical models that predict the most probable continuation of a sequence, not mechanisms that verify the truthfulness of their outputs. This creates a dangerous asymmetry: generating content is instantaneous, while verifying it requires time, expertise, and discipline. In October and November 2025, it was reported[23] that a large consulting firm had delivered two reports to governments containing fabricated or inaccurate citations and references, forcing refunds and corrections and causing significant reputational damage. The reports had been produced with the assistance of generative AI without rigorous verification.

The risk is not in the tool, but in the work processes that do not validate results and sources. A well-intentioned professional can introduce catastrophic errors if they blindly rely on AI outputs without cross-checking them; and this point can only be mitigated with AI awareness and literacy.

23 Fortune (2025a)

AI literacy as a regulatory requirement

Article 4 of the European AI Regulation (AI Act), states[24]: *"Providers and those responsible for the deployment of AI systems shall take measures to ensure that, to the greatest extent possible, their personnel [...] have a sufficient level of AI literacy, taking into account their technical knowledge, experience, education and training, as well as the intended context of use of the AI systems and the persons or groups of persons on whom the AI systems are to be used."*

This is not a recommendation: it is a legal obligation effective February 2, 2025. Organizations that treat this as a compliance checkbox are accumulating operational and reputational risk. Those that approach it as a cultural transformation, integrating continuous learning and communities of practice, are sustainably capturing value[25].

The inevitability of adoption

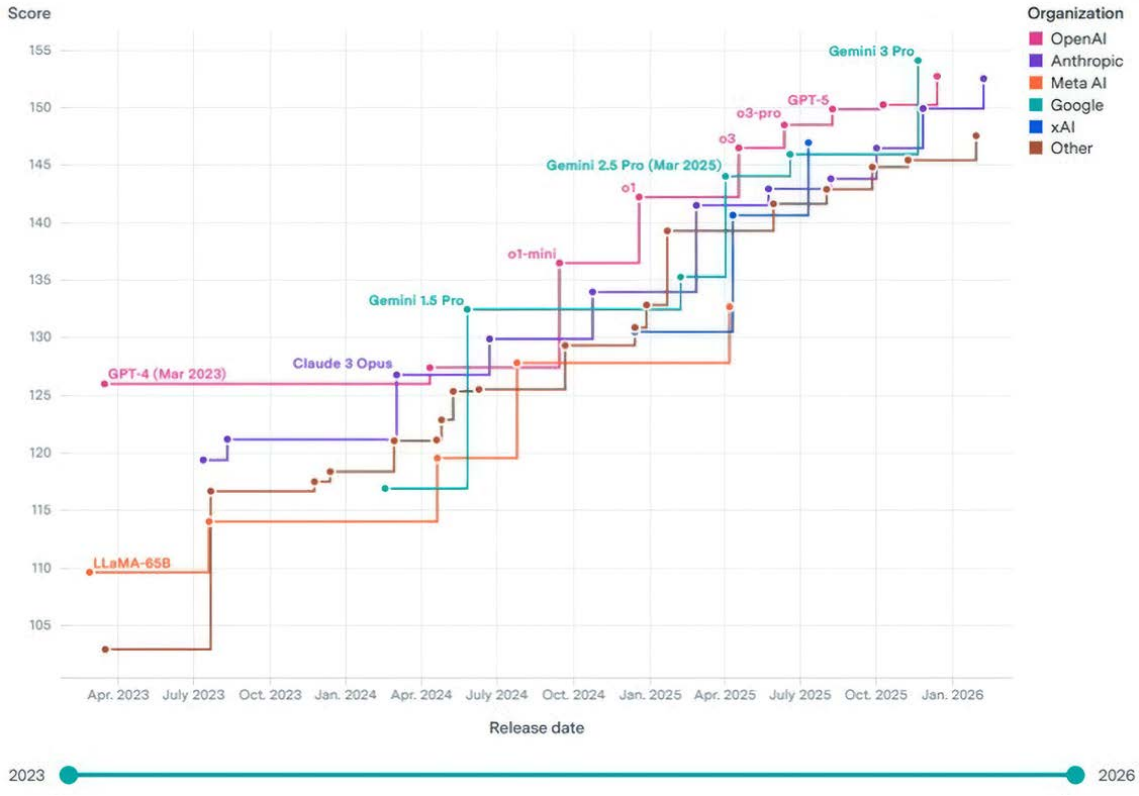
The reality is that this technology is already transforming the way we work, with or without organizational support. If companies do not provide secure corporate tools and adequate training, employees will use uncontrolled alternatives: personal accounts on public platforms, free tools with no privacy guarantees, untraceable systems. Studies[26] indicate that up to 35% of the data professionals upload to unsecured chatbots is confidential. Shadow AI is not a future threat: it is a reality today.

Finally, at the individual level, AI literacy is no longer optional. Professionals who master these tools (i.e., understand when and how to use them, how to verify their outputs, and how to integrate them into complex workflows) will have structural competitive advantages over those who do not. The way of working has changed irreversibly, making adaptation essential for professional and organizational competitiveness.

24 AI Act (2024).
25 Management Solutions (4Q24).

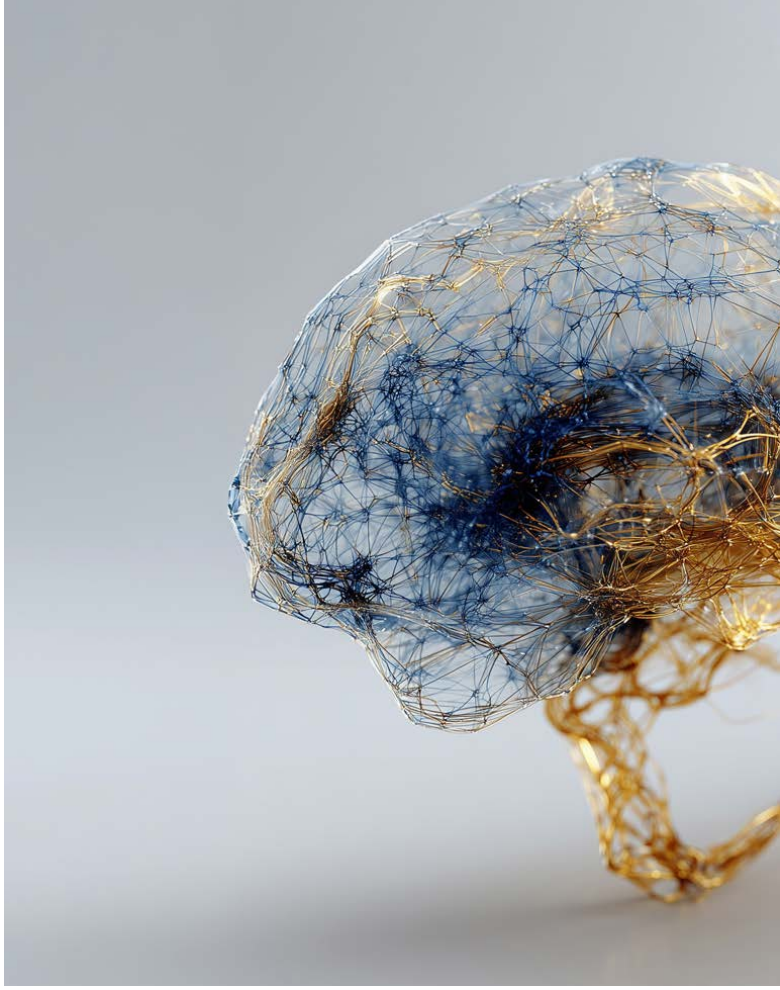
26 Cyberhaven (2025).

Fig. 1. Continuous improvement of LLMs, as measured by a synthetic capabilities index[27].



27 Epoch (2025a).

Machine Learning Accelerated by Generative AI



The persistence of Machine Learning

While generative AI is grabbing headlines, classical Machine Learning (ML) continues to be the backbone of critical applications in sectors such as banking, insurance, retail, logistics, energy and telecommunications. Credit scoring models, fraud detection, demand prediction, inventory optimization, predictive maintenance, customer segmentation and recommendation engines continue to operate using algorithms (logistic regression, random forest, gradient boosting, neural networks...) trained on structured historical data. These systems do not generate content like generative AI: they classify, predict and optimize based on learned patterns.

Generative AI does not replace these models: it makes them faster to develop, easier to document, more efficient to validate and simpler to deploy.

The traditional ML lifecycle: costly and time-consuming

Developing a classic ML model has traditionally been time-intensive for specialized professionals. A typical predictive model requires feature engineering, data preparation and cleaning, algorithm selection and training, rigorous validation, thorough documentation, and deployment on production infrastructure with continuous monitoring. This cycle can extend for months, and each iteration or model update replicates much of the effort.

Generative AI is involved in each of these phases, and has been shown[28] to significantly reduce time requirements, freeing up the most skilled professionals to focus on strategic work. It therefore does not improve them directly in statistical terms; rather, it transforms the work of the teams that build, document, validate and deploy them. The acceleration lies in the human development cycle, not necessarily in the predictive performance of the resulting model.

Accelerating feature engineering

Feature engineering (the process of constructing predictive variables from raw data) is one of the most knowledge-intensive aspects of data science. A data scientist must combine business understanding, statistical intuition, and iterative experimentation to determine which variables are relevant. Generative AI can help accelerate this process:

- Automated variable generation: an LLM can formulate dozens of candidate variables from descriptions of the business problem and the structure of available data[29].
- Translation of business logic to code: an analyst can describe complex logic in natural language (e.g., "I want to capture revenue volatility over the last 12 months adjusted for seasonality") and receive the corresponding SQL or Python code in seconds.
- Data pipeline optimization: Generative AI can review data preparation scripts and identify inefficiencies.

Preliminary studies[30] indicate that generative AI-assisted data scientists "save weeks of manual feature engineering work, improving the performance of various predictive models in multiple business scenarios".

Automated documentation and regulatory compliance

ML models in regulated industries must be thoroughly documented. In banking, for example, supervisors require each regulated model to include documentation covering its purpose, the data used, statistical methodology, validation process, backtesting results, known limitations, and monitoring plan, among other elements. Producing this documentation is tedious and time-consuming for highly skilled profiles, and updates are therefore prone to inconsistencies.[31]

Generative AI can automate much of this process: generate technical documentation from code, translate technical descriptions into summaries understandable to non-technical committees or auditors, and automatically update documentation when a model is retrained with new data.

Assisted validation and bias detection

Validation of ML models is critical and regulatory mandated in many sectors. It includes verifying the statistical robustness of the model, evaluating its behavior in extreme scenarios, and detecting unwanted biases[32]. Generative AI can assist by automatically generating and running complete batteries of relevant statistical tests, evaluating algorithmic fairness using fairness metrics, and proposing and running realistic stress scenarios.

28 Gu (2025).
29 iDanae (2Q23).
30 Ouyang (2025).
31 ECB (2025).
32 iDanae (1Q25).

Industrialized deployment and monitoring

Once validated, the model must be deployed in production environments and continuously monitored for performance degradation (model drift). Generative AI accelerates this phase as well: it can generate infrastructure code (Docker, Kubernetes, CI/CD pipelines) to deploy models in a reproducible and scalable way, produce performance metrics dashboards, set up automatic alerts when anomalies are detected, and generate automatic retraining scripts.

Classical ML is not going away: it is industrialized

Generative AI does not replace traditional Machine Learning, but it does radically transform its lifecycle. Tasks that used to take weeks (feature engineering, documentation, validation) are now solved in days. This does not mean that data scientists are expendable: it means that they can spend more time on strategic work (understanding the business problem, designing innovative model architectures, interpreting results) and less on mechanical tasks (writing repetitive code, writing standard documentation, or running routine tests).

The net result is an acceleration of ML model time-to-market, a reduction in operational costs, and an improvement in the quality and traceability of deployed systems. For organizations that rely heavily on predictive models, this acceleration can translate into significant competitive advantages: the ability to launch customized products faster, to adapt strategies in real time, and to meet regulatory requirements with less operational friction.

Industry note: banking supervisors and ML

For years, European financial institutions avoided using ML in regulated models (particularly in IRB models for regulatory capital calculation) because they believed supervisors would reject them due to explainability concerns. This perception is now changing.[33]

In 2021, the EBA published[34] a discussion paper on ML in IRB models where it recognized that "Machine Learning techniques have the potential to enhance risk differentiation in IRB models" and set out a set of principles-based recommendations to ensure prudent use of ML in this context. The document did not discourage the use of ML; on the contrary, it set out how to make it compatible with existing regulation (CRR)[35].

Practice confirms this: several European institutions have submitted ML-based IRB models (for example, for SME portfolios) to the ECB and have obtained approval, provided they adequately justify explainability through techniques such as SHAP values, sensitivity analysis or decision decompositions. Explainability in ML is not an insurmountable barrier: techniques such as SHAP or LIME allow institutions to justify model decisions to supervisors with sufficient rigour. However, it remains only partially resolved: current XAI methodologies work well for technical and regulatory audiences, but translating those explanations into terms understandable to a retail customer or an executive committee remains an open challenge.

33 Management Solutions (2023).
34 EBA (2021).
35 Management Solutions (3Q23).

Vibe Coding and Augmented Software Development

From traditional programming to cognitive dialog with machines

Software development has historically evolved through leaps of abstraction: from assembly programming to high-level languages, from imperative code to declarative frameworks, from manual development to low-code platforms. Each transition eliminated unnecessary technical complexity and brought software creation closer to human intent.

Generative AI represents a distinct qualitative leap: it turns programming into an iterative conversation with cognitive systems. The programmer no longer writes code line by line: he or she describes the desired behavior in natural language, and the system generates, tests, corrects and documents it. This phenomenon, dubbed "vibe coding" by Andrej Karpathy[36], redefines what it means to program: the developer moves from writing syntax to directing cognitive systems that materialize intent into functional software. In Karpathy's words, "vibe coding is going to terraform software and alter job descriptions"[37].

What is vibe coding really?

Vibe coding is not simply auto-completion of sophisticated code; it is software development through iterative natural language interaction with AI models that maintain memory, context and high-level goal understanding.

Its key components include:

- Semantic problem understanding: the system interprets requirements expressed in natural language and translates them into technical architecture.
- Multi-file code generation: the system produces not just fragments, but complete applications with coherent modular structure.
- Assisted execution, debugging and refactoring: the system executes code, detects errors, proposes corrections and optimizes implementations.
- Automatic production of tests and documentation: the system automatically generates test batteries and technical documentation synchronized with the code.

The difference from traditional code assistants is fundamental: traditional tools complete lines or functions, whereas vibe coding systems understand objectives, maintain architectural coherence across extended sessions, and act as cognitive collaborators rather than passive tools.

Acceleration and Democratization of Software Development

The impact of AI-assisted coding on development speed is both measurable and substantial. A field study[38] with 4,867 developers found that task completion rate increased by 26%. An experiment[39] with 187,489 developers showed that they spent 12.4% more time on core programming activities, while reducing time spent on project management and administrative tasks by 24.9%. In other words: projects that used to take months are now completed in weeks or days.

Another disruptive effect is the equalizer: AI narrows the productivity gap between junior and senior developers. While junior developers experience[40] productivity gains of 21% to 40%, senior developers improve by 7% to 16%. This does not mean experience no longer matters; rather, the development bottleneck is shifting. Success now depends less on mastering syntax and more on understanding problems, designing robust architectures, and formulating constraints accurately.

This shift in the skills gap has a direct organizational consequence for those that act on it: the profound democratization of software creation. Business analysts, product managers, consultants and scientists are generating functional prototypes without relying on engineering teams as intermediaries. Software is no longer the exclusive domain of specialized technical profiles. The barrier to entry has moved from knowledge of programming languages to the ability to grasp problems, define objectives, and specify constraints clearly.

The net result is a structural reduction in the marginal cost of creating software, leading to a change in the economics of technology production.

Transformation of technology teams

Within technology organizations, the distribution of roles is changing rapidly. Teams need fewer "low-level programmers" writing routine code, and more system architects, solution designers, quality validators and technical risk auditors. The role of senior developers is evolving: they spend less time on syntax and more on governing architecture, security, scalability and technical debt management.

Research shows[41] that AI-assisted teams require 79.3% fewer contributors per project on average, without sacrificing technical complexity. Small teams are producing systems of a scale and sophistication previously reserved for large engineering departments. In addition, exploration of new technologies is up 21.8%, suggesting that developers are freeing up cognitive capacity to learn, experiment, and expand their technical capabilities.

36 Andrej Karpathy (b. 1986), former Director of AI at Tesla and former senior researcher at OpenAI, with key contributions in deep learning, computer vision and autonomous systems.
37 Business Insider (2025b).
38 Stanford (2025).
39 Ibíd.
40 Ibíd.
41 Ibíd.

Quality, technical debt and risk

However, speed has hidden costs. Generating code is instantaneous; maintaining, debugging and scaling it remains difficult. Vibe coding introduces new risk vectors:

- Hidden fragility: generated code may work superficially, but it can contain inefficiencies, vulnerabilities, or unstable dependencies that only surface under production load or in extreme cases. To this we must add an earlier risk in the chain: ambiguous or poorly formulated specifications which, in the past, a senior developer would have challenged or sensibly reinterpreted are now executed literally, without any corrective friction, silently propagating the error from the requirement through to the final product.
- Model dependency: if the AI system that generated the code disappears or changes significantly, the ability to maintain or extend the software is degraded.
- Systemic bugs replicated at scale: a bug in the prompt or model logic can instantly propagate to dozens of projects, multiplying the impact of bugs that previously would have been isolated.

The nature of technical debt is also changing. Before, technical debt was primarily code debt: fast implementations, postponed refactoring, duplication of logic. Now, technical debt includes architecture debt (design decisions implicit in interactions with the AI), prompts debt (poorly formulated instructions that generate suboptimal but functional code), and traceability debt (loss of understanding of why the code does what it does).

Software governance in the age of vibe coding

If anyone can create software through AI conversation, organizational risk multiplies. The problem is no longer just what code exists, but what cognitive system produced it, under what instructions and with what degree of autonomy.

Leading security and development frameworks are already warning of this change. OWASP identifies[42] new structural risks in LLM-based applications, such as prompt injection, insecure output handling and excessive agency: giving an AI system the ability to act without sufficient controls.

At the same time, NIST insists[43] that the classic principles of secure development (traceability, review, testing, change control, continuous validation) must also apply to AI-generated content and the mechanisms that turn it into executable changes.

Consequently, software governance ceases to be solely code governance and becomes cognitive systems governance, forcing the introduction of new layers of control:

- Prompt repositories: catalog, version and audit instructions that generate critical code, as the prompt becomes a risk surface.
- Intention version management: record not only what code was produced, but what objective was pursued and what restrictions were defined.
- Autonomy and permissions control: explicitly limit what actions AI can perform (repository modification, deployments, commands, data access).
- Traceability of design decisions: document which decisions were human, and which were proposed or executed by the AI.
- Validation and assisted review: conduct systematic review of diffs, automatic tests and behavioral audits before accepting changes in production.

Agent-based development tools themselves already explicitly recommend these guardrails (change review, permission control and caution with automatic execution), reflecting that the risk is not theoretical: the attack and failure surface has shifted from isolated code to the full intent → generation → execution loop.

Strategic implications

Vibe coding is not just a technology trend; it is a macroeconomic variable. Organizations that master this capability operate with structural competitive advantages: extremely compressed time-to-market, massive low-cost experimentation, and accelerated organizational adaptability.

For traditional companies, the implication is clear: they are competing against organizations that iterate ten times faster, with teams ten times smaller, and with marginal costs of development that tend to zero. The speed of software creation goes from being an internal operational metric to a determinant of competitive survival.

42 OWASP (2025).
43 NIST (2024a).

Agentic AI and Autonomous Systems

From conversational assistants to autonomous operators

Generative AI has transformed intellectual work by enabling professionals to generate content, analyze information and obtain answers through natural conversation. However, these systems remain fundamentally reactive: they respond to prompts, but do not act independently on real systems. Agentic AI represents a qualitative leap: systems that plan, execute complex tasks, make decisions within defined boundaries and operate on corporate infrastructures with full traceability.

An agentic system operates through autonomous agents, each with specific capabilities (reasoning, memory, tool usage, planning), that collaborate to achieve a defined goal. The incremental capabilities of agentic AI over generative AI are structural[44]:

- **State and memory:** maintains persistent context between interactions, not just within an isolated conversation.
- **Dynamic planning:** decomposes complex objectives into subtasks, prioritizes them and replans according to results.
- **Execution on real systems:** uses tools and APIs to modify data, execute commands and complete transactions.
- **Multi-agent orchestration:** coordinates specialized agents under a central coordinator that manages dependencies and information transfer.
- **Full traceability:** produces logs, evidence and justifications of each executed step, enabling auditing and supervision.

Agentic AI in production

Agentic AI operates today in global organizations that manage millions of daily transactions. To cite a few examples:

- **Deutsche Bank:** is deploying an AI voice-enabled agent ("AI banking butler") that acts as a proactive conversational agent and covers everything from care and support to transaction execution and advisory services[45]. The program is part of a technology investment of approximately 600 million euros, with a recurring savings target of around 300 million euros per year by 2028, and is expected to result in a workforce reduction of roughly 10%.[46]
- **Ryt Bank:** Malaysian bank that advertises itself as "AI-first", operates on an architecture of specialized agents where the customer interacts in natural language and the agents interpret the intention, orchestrate processes and execute real transactions on the banking core. The system handles around 80,000 transactions per month, has reduced processes that required 5-8 screens to a single conversational interaction, and has shown significant improvement in customer retention.[47]

44 iDanae (2Q25).
45 Deutsche Bank (2025).
46 Financial News London (2025).
47 Ryt Bank (2025).

- **Walmart:** in its automated distribution centers, autonomous decision systems coordinate real-time stocking, replenishment and order picking. Walmart reports that these centers double processing capacity with approximately half the staff versus traditional centers, evidencing a structural shift in logistics productivity[48].
- **Amazon:** its new logistics management system, Sequoia, orchestrates robots, inventory and order flows through autonomous software that decides where to store, when to move inventory and how to stock pick stations. Amazon reports reductions of up to 75% in inventory handling time and up to 25% in order processing time at centers where is deployed.[49]
- **DHL:** in picking operations with autonomous robots integrated with the WMS, systems automatically assign work, optimize routes and close tasks. In productive deployments, DHL has reported productivity increases of up to 180% or more in units per hour, along with significant improvements in quality and accuracy.[50]

48 Business Insider (2025a).
49 Amazon (2023).
50 DHL (2024).

But what is an agentic system? Five modular layers

These functional capabilities materialize in a specific modular architecture (not simply a language model with tool access) composed of five interdependent layers:

1. **Interface and perception:** receives user goals, delivers end results, and perceives events from the environment via API gateways, endpoints, and input connectors.
2. **Orchestration and scheduling:** decomposes complex goals into manageable tasks, decides which agent executes each task and in what order, and manages priority queuing and routing.
3. **Agent core:** autonomous workers, each with a specific role, cognitive core (LLM) and ReAct (Reason + Act) control loop that combines reasoning and iterative action.
4. **Tools and services:** library of external capabilities (search, code generation, corporate APIs) with standardized connectivity through protocols such as MCP and prompts management.
5. **Memory and knowledge:** stores short-term information (conversation history), long-term information (vector database with past experiences) and corporate knowledge base.

This architecture turns autonomy into something traceable, auditable and governable. Every decision, every action, every invocation of tools is recorded, enabling effective human oversight and regulatory compliance.

MCP: the missing link to scalability

Agentic architecture faces a fundamental technical challenge: the exponential complexity of integrations. Traditionally, each language model requires a proprietary integration with each tool. Changing models requires rewriting all integrations, and adding a new tool requires integrating it with all existing models. The result is exponential technical debt.

Model Context Protocol (MCP) solves this problem through a universal abstraction layer. MCP is an open protocol that standardizes how AI models interact with external applications, data sources and tools. MCP servers expose capabilities, such as resources, prompts, and tools, that MCP clients (agents or models) can consume on demand. Once a tool is connected to MCP, it becomes immediately accessible to any current or future agent, without the need to redevelop integrations.

The impact of MCP is transformative: it moves from hand-crafted integrations to universally reusable assets, facilitates the transition from isolated prototypes to scalable ecosystems, enables agents to acquire new capabilities without redeployments, and dramatically reduces the cost of maintenance and evolution. Without MCP or equivalent standards, agentic AI at enterprise scale is not technically sustainable.

Governance and control: the real challenge

Building agents is relatively straightforward with current frameworks; governing them at enterprise scale is the real challenge. Sustainable adoption requires balancing four capabilities:

- **Technology and development:** multi-agent orchestration with effective coordination, operational and long-term memory management, modular architectures that allow maintenance and evolution, and secure integration with corporate systems through standards such as MCP.
- **Continuous evaluation:** performance, quality and efficiency metrics, monitoring of deviations and anomalies, cost control (calls to models multiply exponentially) and full traceability of actions and decisions.
- **Governance and compliance:** human oversight through approval and escalation mechanisms, (bearing in mind that human supervisory capacity has a ceiling; once that limit is exceeded, supervision becomes nominal and creates a false sense of control), explicit controls and limits on what actions each agent can execute, transparency and functional explainability of decisions, and compliance with the AI Act and internal risk policies.
- **Industrialization and operating model:** 24/7 operation with continuous maintenance, CI/CD pipelines to securely deploy and update agents, active cost management in production, and resilience to failures or unexpected behaviors.

Strategic implications

The adoption of agentic AI has three critical strategic implications:

- **Transformation of the operating model:** agentic AI introduces a structural change where human teams collaborate with autonomous agents operating 24/7 without continuous supervision. It radically compresses time-to-market: tasks that used to take weeks are now solved in days through delegation to specialized agents.
- **Risk of cost escalation:** unlike traditional generative AI, agents multiply calls to models and tools. A viable prototype can become an economically unsustainable system if it is not designed with cost controls from the start.
- **Industrialization as a barrier to entry:** building agentic prototypes is possible with current frameworks, but operating, scaling, maintaining, securing and auditing them in production requires organizational capabilities that most companies have not yet developed. The competitive advantage is not in technology, but in the ability to industrialize it with effective governance.

AI in Robotics and Physical Systems

From digital to real-world action

For years, industrial robotics has operated through systems programmed for repetitive tasks in highly controlled environments: robotic arms that assemble components following fixed sequences, automated guided vehicles (AGVs) that follow predefined routes, or pick-and-place systems that recognize objects in exact positions. These robots execute precise movements but lack adaptability: any change in the environment (a misplaced object, a variation in texture, an unexpected obstacle) requires reprogramming or human intervention.

The integration of generative AI, advanced vision models and reinforcement learning is transforming this reality in industry: in 2023 alone, more than 276,000 industrial robots were installed in China (Fig. 2), which already accounts for more than half of the world's installations, and the proportion of collaborative robots has quadrupled in six years[51]. Today's robots perceive their environment through real-time computer vision, interpret instructions in natural language, plan complex sequences of actions, adapt to unforeseen changes without reprogramming, and learn from each interaction to continuously improve. AI turns rigid industrial machines into autonomous systems capable of operating in unstructured environments and performing tasks that previously required human intelligence.

51 Stanford (2025).

Humanoid robots: from the lab to the factory floor

Humanoid robotics has taken a quantum leap in the last two years, moving from spectacular lab demonstrations to real industrial deployments.

Figure AI, a startup valued at approximately $39 billion, completed an 11-month deployment of its Figure 02 robots at BMW's Spartanburg (South Carolina) plant in 2025[52]. The two humanoid robots worked 10-hour shifts Monday through Friday, accumulating 1,250 hours of operation, loading more than 90,000 sheet metal parts to 5-millimeter tolerances in 2 seconds per part, and contributing to the production of more than 30,000 BMW X3 vehicles.

Figure has launched its third generation, the Figure 03, designed specifically for volume production. The company built BotQ, a manufacturing facility dedicated to humanoid robots with an initial capacity of 12,000 units per year and a goal of producing 100,000 robots in four years. The Figure 03 incorporates inductive wireless charging (2 kW via foot coils that allow the robot to simply step onto a base to recharge), a redesigned vision system with double the refresh rate and one-quarter the latency of the previous generation, and integrated palm cameras for redundant visual feedback during fine manipulation. The company projects that these robots will operate using its proprietary Helix AI, massively trained with teleoperation data and human demonstrations.

For its part, Tesla is aggressively ramping up production of its Optimus humanoid robot. The company announced plans to build a production line capable of manufacturing one million units annually, with startup expected toward the end of 2026[53]. Elon Musk stated in October 2025 that Optimus version 3 will have "hands that are an incredible piece of engineering" with full human range of motion (22 degrees of freedom) and that the robot will be so realistic that "you'll need to touch it to believe it's really a robot"[54]. Tesla produced several thousand units in 2025 for internal use in its factories (primarily battery and component handling tasks), and plans to scale to 50,000-100,000 units in 2026[55]. Musk estimates[56] that, at volumes above 1 million units per year, Optimus' production costs will drop below $20,000, roughly half the cost of an equivalently scaled Model Y. The retail price, however, will be significantly higher and determined by market demand.

Boston Dynamics, a historic reference in dynamic robotics, retired its hydraulic Atlas (famous for backflips and parkour) in April 2024 and launched an all-electric Atlas designed for real industrial applications[57]. The new Atlas integrates high-performance custom actuators with range of motion that exceeds human capabilities: its head and torso can rotate 180 degrees independently, its joints have extreme flexibility, and it is designed to exploit its own mechanical anatomy, not to be limited to human postures, although many of its control capabilities are trained from human movement. Boston Dynamics emphasizes that Atlas will prioritize speed and efficiency over anthropomorphic appearance.

52 Figure (2025).
53 Teslarati (2025).
54 Tesla Car World (2025).
55 Fortune (2025b).
56 Notateslaapp (2025).
57 Boston Dynamics (2025a).

In August 2025, Boston Dynamics and Toyota Research Institute demonstrated[58] Atlas operating through Large Behavior Models (LBMs): end-to-end policies trained with extensive demonstrations and language annotations that coordinate locomotion and manipulation simultaneously. A single behavior model directly controls the entire robot, treating hands and feet almost identically, without separating low-level locomotion control from manipulation control. In public videos[59] , Atlas performs continuous sequences of sorting and packing tasks in simulated factory environments, reacting autonomously to unexpected physical disturbances (such as researchers closing boxes or pushing objects) without interrupting the task. Boston Dynamics plans pilots[60] with Hyundai in 2026 and limited commercial production starting in 2027.

Beyond manufacturing and logistics, humanoid robotics is opening a second front of impact: the care of older, dependent or disabled individuals. In economies facing accelerated structural ageing, this application has strategic relevance comparable to that of industrial use, with its own ethical and regulatory implications that existing frameworks have only just begun to address.

Strategic implications and operational challenges

The integration of AI into humanoid robotics introduces structural changes in manufacturing and logistics. Productivity increases radically: robots operate 24/7 without fatigue, breaks or performance variation. Recurring operational costs (energy, preventive maintenance) are predictable and decreasing with economies of scale, although the initial investment remains significant.

However, critical challenges arise. The impact on employment is real and concentrated: repetitive manual tasks in manufacturing, assembly and material handling face accelerated automation. Organizations adopting humanoid robotics must manage job transitions, reskilling programs, and increasing regulatory and societal expectations about corporate responsibility, challenges shared with the broader adoption of Artificial Intelligence more, but which here carry greater sectoral concentration and social visibility.

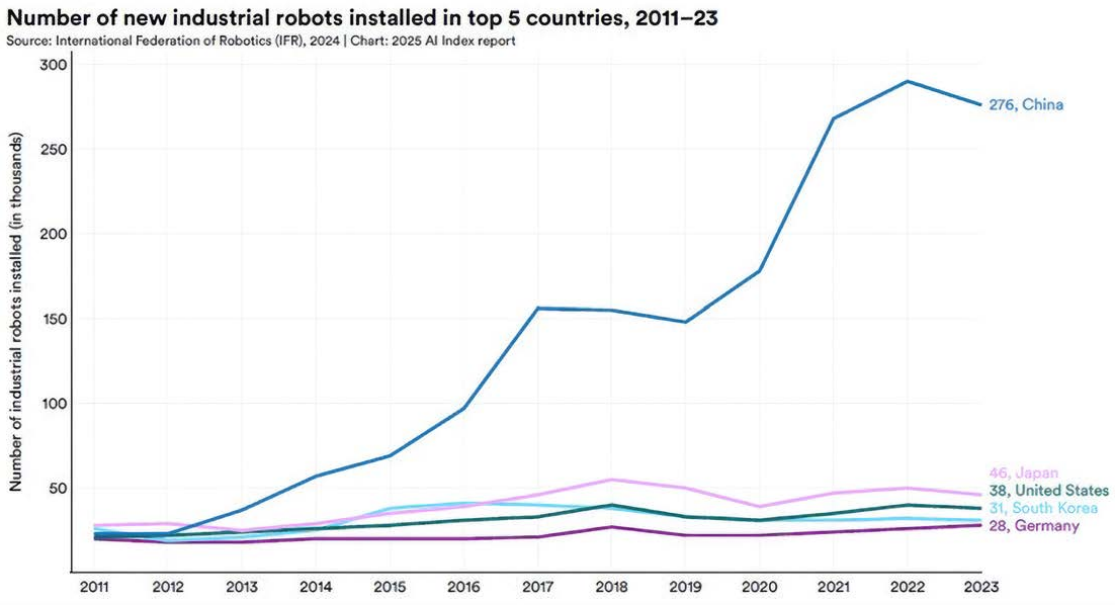
Vendor dependence intensifies: companies adopting robots are tied to their proprietary ecosystems of hardware, software, AI upgrades, and technical support. Technological obsolescence is rapid: a robot purchased today can be surpassed in capabilities by the next generation in two years, raising questions about investment cycles and upgrade strategies.

And new operational risks emerge: autonomous systems operating in physical environments shared with humans can cause injury, property damage or critical operational disruptions if they fail. Robotics with AI requires robust safety frameworks (redundant sensors, emergency shutdown systems, dynamic exclusion zones), certified fail-safe protocols, and effective human supervision even in nominally autonomous operations.

AI in robotics is not a future trend, it is an operational reality undergoing accelerated industrialization. Organizations that strategically evaluate when and where to adopt humanoid robotics (not all tasks justify the investment) will capture sustainable competitive advantages.

58 Boston Dynamics (2025a).
59 Toyota Research Institute (2025).
60 Hyundai (2025).

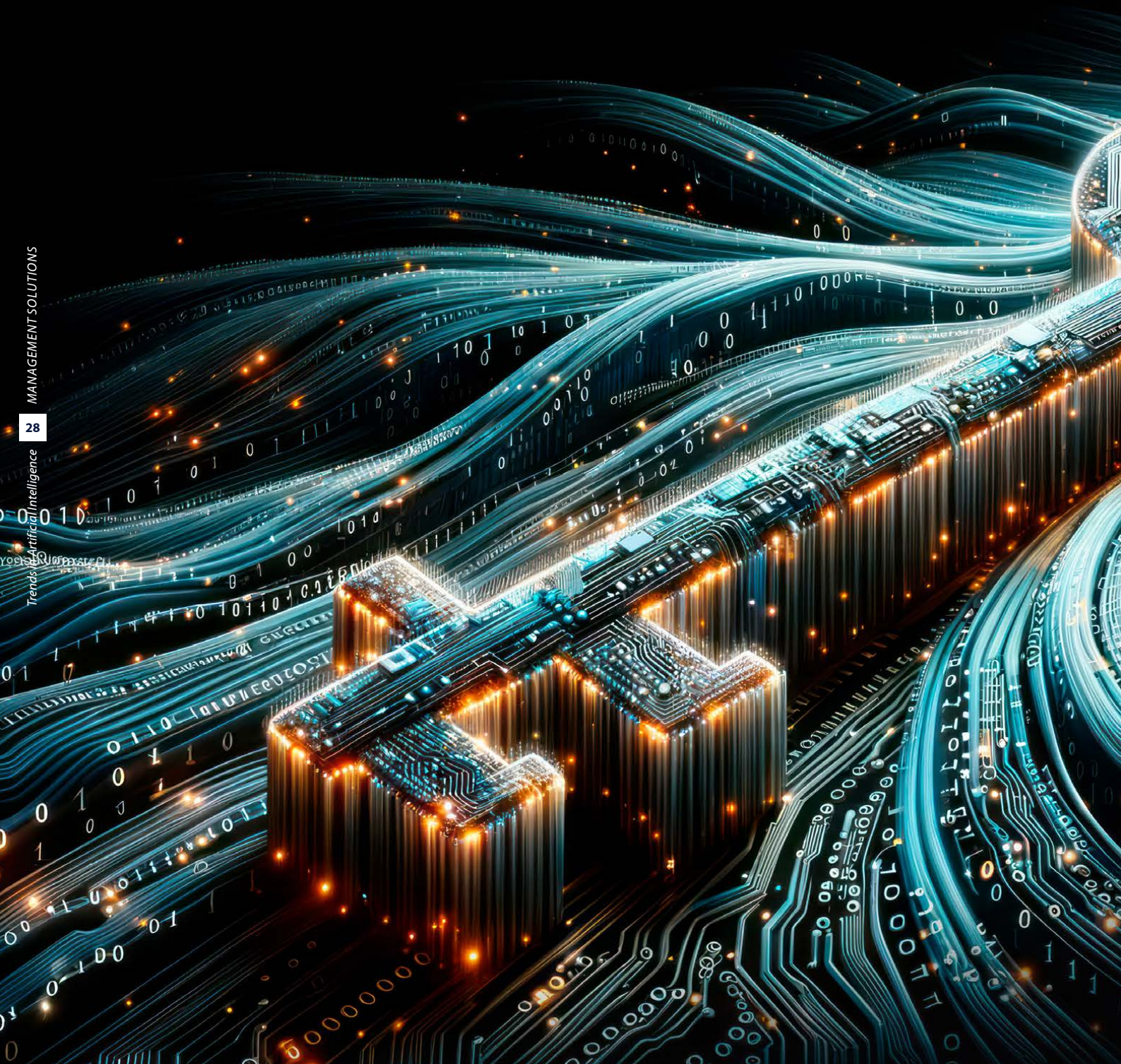
Fig. 2. New industrial robots installed. Source: Stanford (2025).



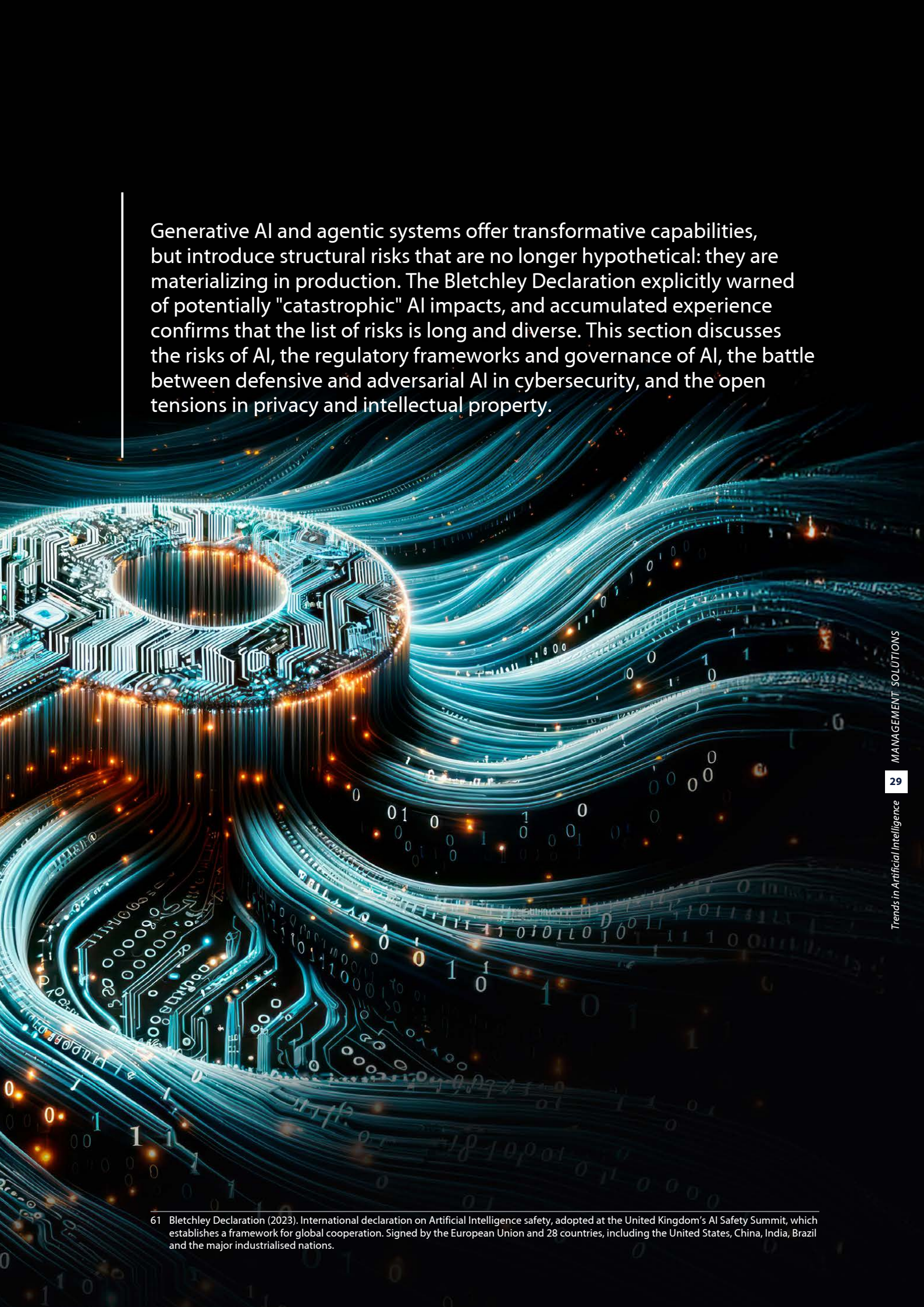
04 | AI Risks, Regulation and Safety

«There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models».

Bletchley Declaration[61]



Generative AI and agentic systems offer transformative capabilities, but introduce structural risks that are no longer hypothetical: they are materializing in production. The Bletchley Declaration explicitly warned of potentially "catastrophic" AI impacts, and accumulated experience confirms that the list of risks is long and diverse. This section discusses the risks of AI, the regulatory frameworks and governance of AI, the battle between defensive and adversarial AI in cybersecurity, and the open tensions in privacy and intellectual property.



61 Bletchley Declaration (2023). International declaration on Artificial Intelligence safety, adopted at the United Kingdom's AI Safety Summit, which establishes a framework for global cooperation. Signed by the European Union and 28 countries, including the United States, China, India, Brazil and the major industrialised nations.

AI Risks

AI does not create risk: it amplifies it

The adoption of AI does not introduce fundamentally new risks, though there are some exceptions. Instead, it drastically amplifies existing risks: operational, model, technological, vendor, legal, reputational, compliance, strategic, social, etc. The difference lies not in the nature of the risk, but in its speed of propagation, scale of impact, and difficulty of containment.

Except for a few emerging categories (such as toxic content generation, certain forms of cognitive manipulation through deepfakes, or prompt injection attacks), most of the risks associated with AI are accelerated, automated and massive versions of known problems. An algorithmic bias is, in essence, a human bias systematized and replicated millions of times. An information leak due to misuse of a chatbot is ultimately an information leak.

Four interconnected dimensions

In practice, these risks manifest across four major, interconnected dimensions (Fig. 3):

1. Security and compliance

AI introduces new attack surfaces and complicates regulatory compliance. Privacy and information security face specific threats: unintentional leaks of confidential data from AI systems, emerging technical vulnerabilities such as prompt injection (malicious instructions embedded in inputs that "trick" the model into ignoring restrictions) or jailbreaks (techniques for circumventing security controls), and accidental exposure of sensitive information when professionals use unsecured tools.

Intellectual property becomes ambiguous: who owns the AI-generated code, what happens when a model reproduces copyrighted fragments? As an example, the New York Times has open litigation with OpenAI/Microsoft for using copyrighted content without a license, among many other ongoing lawsuits.[62]

Traceability and reproducibility are degraded when critical decisions rely on models that continually evolve through automatic retraining, making it difficult to reconstruct exactly which version of the system produced which result at which time.

And regulatory non-compliance now has direct and visible financial consequences: the European AI Act[63] provides for fines of up to 35 million euros or 7% of annual global turnover, making AI compliance a material risk of the first order, higher even than data protection.[64]

2. Quality, reliability, lock-in and costs

As already mentioned, Generative AI is intended to produce plausible outputs, not necessarily correct ones. Hallucinations (generation of false information presented with confidence) are not occasional bugs, but behaviors intrinsic to the actual design of the models. Non-deterministic behavior implies that the same input can produce different outputs in the same model, which complicates validation, auditing and certification of critical processes.

Silent model drift represents one of the most insidious risks: a model that was working properly can degrade in a short time because the data distribution changed, without the system generating early warnings. A model can continue to operate with the appearance of normality until someone manually detects anomalies in the results.

Over-reliance on AI in critical tasks creates operational fragility: if the AI system fails, crashes, or becomes prohibitively expensive, can the organization continue to operate? Are there human contingency procedures? And loss of effective human control occurs when decisions are delegated to systems whose internal reasoning is opaque even to those who develop and operate them.

Moreover, organizations build structural dependencies on a few model providers (OpenAI, Anthropic, Google) and infrastructure (AWS, Azure, GCP). Changes in pricing, terms of service or operational outages can cripple critical processes simultaneously. Migrating between ecosystems involves rewriting integrations, retraining workflows, recertifying compliance and assuming prohibitive costs, which is not always feasible; diversifying suppliers, although costly, is the structural response to this risk.

Added to the above risks is an economic dimension that traditional management frameworks underestimate. Unit inference costs are structurally declining (approximately 10x every 12 months), but this decline does not shield against the exponential growth in volume: the agentic systems have cost structures with non-linear scalability: each agent multiplies calls to models and tools, and without real-time token monitoring and explicit spending limits built into the design, a viable prototype can become an unsustainable system before anyone detects it.[65] Added to this is uncertainty about return on investment: there is no definitive answer as to whether heavy investment in AI will produce the expected benefits, and many organizations are moving forward driven by competitive pressure rather than a robust business case. Gartner synthesizes both risks in one prediction: more than 40% of agentic projects will be cancelled by 2027 due to cost escalation, unclear business value, and inadequate risk controls.[66] Lock-in exacerbates the problem structurally: dependence on a single vendor eliminates the ability to migrate to more efficient architectures when they become available.

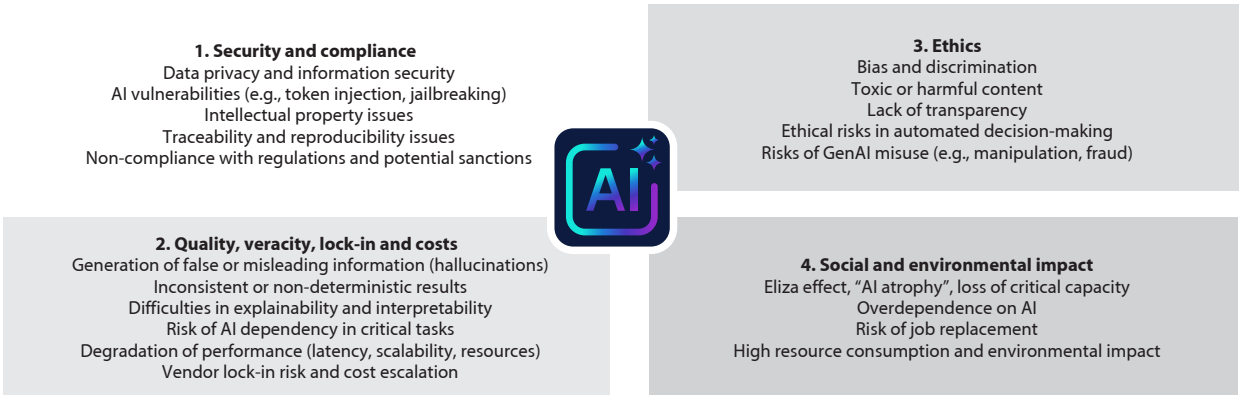
3. Ethics and Automated Decision Making

Algorithmic biases amplify pre-existing biases in historical data: if a personnel selection model is trained with data from an organization historically biased towards certain profiles, the model

62 Brown (2025).
63 AI Act (2024).
64 Management Solutions (4Q24).

65 The estimation of costs before development and continuous monitoring in production are two capabilities that organizations are beginning to formalize within their AI governance frameworks. GenMS Tracker, a proprietary solution by Management Solutions, addresses both dimensions: it allows estimating the development and operational cost of an AI system based on a natural-language description, and it monitors real-time consumption in production with alerts in case of deviations from the defined budget.
66 Gartner (2025b).

Fig. 3. Some risks of generative AI.



will systematize and scale that bias[67] . The lack of transparency and the challenges of explainability make it harder to meet regulatory requirements that expect automated decisions to be understandable and justifiable, especially when they affect fundamental rights.

Automated decisions with direct human impact (credit approval, medical diagnoses, criminal risk assessment, candidate selection) raise liability issues: who is liable when the model makes a mistake with serious consequences: the provider of the base model, the team that customized it, the user who wrote the prompt, the committee that approved its deployment? The chain of responsibility becomes complex, creating accountability gaps that neither regulation nor practice has fully resolved.

4. Social and organizational impact

Job transformation is no longer speculative: routine manual and cognitive tasks face accelerating automation, and organizations must manage job transitions, reskilling programs, and increasing regulatory and social expectations about corporate responsibility.

Cognitive dependence and erosion of critical capabilities represent a major risk: if an entire generation of professionals comes to rely exclusively with AI as an intermediary, will they retain sufficient intuition and knowledge to critically assess outcomes? What happens if AI is not available? The pressure on resources and environmental footprint is significant: training advanced models consumes as much energy as thousands of homes would use over several months, and the continuous, large-scale inference performed by these models further increases energy demand, raising questions about long-term sustainability.[68]

Nonlinear amplification

AI introduces a key phenomenon in risk management: nonlinear amplification. A minor glitch (a data bias, a poorly designed prompt, a permission misconfiguration) can escalate in minutes and simultaneously affect processes, customers, regulators and reputation. The greater the autonomy and integration of AI into critical processes, the wider the potential impact of any failure.

A concrete example: a customer service model that, after a minor change in the system prompt, starts to reveal confidential information from other customers in "as few as" 0.01% of conversations, which is almost undetectable in testing. In a system that handles 100,000 interactions per day, this means 10 leaks per day before the pattern is detected. By the time the problem is identified, hundreds of incidents have already occurred, each with regulatory, contractual and reputational implications.

AI in the corporate risk taxonomy

Organizations are taking one of three approaches to integrating AI risk into their corporate risk framework:

- Treat it as a top-level risk, creating a new category in the taxonomy (a very rare approach, but useful for gaining executive visibility in the early stages of mass adoption).
- Treat it as a second level risk, linked to technological, model or operational risks within the existing taxonomy.
- Do not treat it as an independent category, but as a transversal driver that amplifies existing risks (model risk, supplier risk, reputational risk, etc. – amplified by AI).

In practice, the label is less important than the organization's ability to identify, prevent, control and mitigate these risks systematically and continuously, with clear metrics, assigned responsibilities and defined escalation mechanisms.

Strategic implication

AI risk management has become a strategic decision that conditions the speed of adoption, operational viability and institutional credibility of the organization.

Organizations that approach AI primarily as a technological innovation project encounter friction with the regulatory framework, operational incidents and reputational tensions. By integrating AI from the outset into governance, internal controls, and risk management frameworks, organizations can scale more smoothly, achieve greater stability, reduce friction, and build greater trust from regulators, the market, and their own professionals.

Sustainable competitive advantage arises from a superior ability to govern AI: well-defined control frameworks, explicit responsibilities, continuous monitoring and an organizational culture that keeps human judgment at the center of critical decisions.

67 Management Solutions (2023).
68 iDanae (1Q24).

AI Regulation, Oversight and Standards

AI, a regulated activity

The rapid adoption of AI and the emergence of its associated risks have prompted an unprecedented global regulatory response. Unlike previous technology cycles, where regulation reacted with years of delay, AI is being regulated in parallel with its massive deployment, precisely because its systemic risks are already materializing.

Europe has taken the regulatory lead with the AI Act[69] , the first comprehensive legal framework on AI. It is followed by significant initiatives in the US, China, the UK, Canada and other countries, along with a growing ecosystem of technical standards and voluntary frameworks that seek to operationalize principles of governance, security and ethics.[70]

The consequence for organizations is clear: AI management is no longer a technological issue but a structural regulatory domain, comparable in impact to that of data protection, financial markets or operational security.

The AI Act: the most prescriptive regulation

The European AI Regulation (Regulation (EU) 2024/1689) establishes a risk-based regulatory model, structuring obligations according to the potential impact of systems on fundamental rights, security and public order.

The framework classifies AI systems into four main categories:

1. Unacceptable risk: prohibited systems (cognitive manipulation, social scoring, certain forms of biometric surveillance).
2. High risk: systems used in critical areas such as credit, employment, education, infrastructure, justice, healthcare, and biometrics. This is the core of the AI Act.
3. Limited risk: systems requiring transparency obligations.
4. Minimal risk: systems that are free to use under some general principles.

For high-risk systems, the AI Act introduces a set of structural obligations:

- Documented risk management system
- Data governance and quality control
- Comprehensive technical documentation
- Logging and traceability
- Effective human oversight
- Accuracy, robustness and cybersecurity requirements
- Pre-deployment compliance assessment and ongoing post-market surveillance.

Penalties for serious non-compliance with the AI Act reach €35 million or 7% of annual global turnover, exceeding even the GDPR regime (which reaches 4%).

Supervision in Europe: from innovation to permanent control

The AI Act establishes a new institutional architecture:

- AI Office: central technical body of the European Commission for AI, especially GPAI.
- National supervisory authorities: supervise and sanction compliance with the AI Act in each country.
- AI Board: forum for coordination and common interpretation between the Commission and the national authorities.
- Cross-border cooperation mechanisms: rules for coordinating cases and investigations involving several Member States.

Supervision will not be limited to one-off audits: a continuous surveillance model is established, with reporting obligations, serious incident management, withdrawal of unsafe systems and inspection powers comparable to those of financial regulators.

In practice, the AI Act's supervisory architecture is still under construction. The IA Board, the central coordinating body between Member States and the Commission, held its first formal meeting[71] in September 2024, and has focused on organizational issues, codes of practice for AIFM models, and coordination of national authorities. The minutes reveal a process still in the organizational phase: selection of a chair, creation of subgroups and discussion on priority deliverables.

For their part, the national supervisory authorities have hardly been designated. Spain created the AESIA (Agencia Española de Supervisión de la Inteligencia Artificial)[72] in August 2023, becoming the first AI authority in Europe, and started effective operations in February 2025 with oversight functions for prohibited systems. Other member states are in earlier stages of designating authorities.

69 AI Act (2024).
70 Management Solutions (4Q24a).
71 AI Board (2026).
72 AESIA (2026).

The standards ecosystem: from principles to operation

In parallel to regulation, a web of technical and management standards that seek to establish norms for AI operational practice is consolidating; among them:

- ISO/IEC 42001: establishes requirements for an AI management system (AIMS), ISO 27001/9001-style, for governance and responsible use of AI.
- ISO/IEC 23894: provides a detailed AI risk management lifecycle framework, intended to integrate with enterprise risk management frameworks.
- ISO/IEC 5259: AI data quality framework and metrics.
- NIST AI Risk Management Framework: a framework for AI risk management, organized around the Govern–Map–Measure–Manage functions and designed to support trustworthy AI principles.
- OECD AI Principles: high-level principles (human values, transparency, robustness, accountability, inclusiveness) that have directly influenced the AI Act and other regulatory frameworks.

These standards play an essential role: they establish concrete controls, auditable processes, and verifiable metrics. For many organizations, they are becoming the technical foundation of their compliance programs.

Geopolitical fragmentation and operational complexity

The European risk-based and binding approach is not being replicated globally:

- **The United States** lacks a federal framework, relying instead on a fragmented sectoral approach by agencies (FTC, FDA, EEOC), with no equivalent to the AI Act. This is supplemented by executive orders, agency guidance, and regulation in some states (e.g., California, Colorado, Connecticut, Utah).[73]
- **China** articulates AI within a state strategy of "digital sovereignty", with specific rules on algorithms, deepfakes and generative models, strong data control and compulsory licensing for certain high-impact systems.[74]
- **UK** maintains a pro-innovation approach without a horizontal AI law, relying on sectoral authorities and common AI regulatory principles, supported by guidelines and regulatory sandboxes.[75]
- **Brazil** passed a Senate bill in December 2024 structurally similar to the AI Act, featuring a risk-based approach (prohibited/high/limited/minimal), specific obligations for high-risk systems, fines up to R$50 million or 2% of turnover. The bill is pending approval in the Chamber of Deputies, with entry into force expected one year after enactment.[76]
- **Mexico** lacks specific AI regulation: more than 60 bills have been submitted since 2020 without approval. In February 2025 a constitutional reform was introduced to grant the federal government competence over AI, which should lead to a General Law. Until then, there is no specific legal framework, only voluntary guidelines aligned with UNESCO and OECD principles.[77]
- **Australia** maintains a voluntary and principles-based approach: AI Ethics Principles (2019, eight voluntary principles) and Guidance for AI Adoption (October 2025, replacing the Voluntary AI Safety Standard of 2024). There is no binding legislation specific to AI, and the government has instead focused on reinforcing existing laws (privacy, consumer protection, sectoral).[78]

Strategic implications

Studies agree[79] that the divergence between these models (EU more prescriptive, US fragmented and sectoral, China highly centralized, UK and Australia more principle-based) forces global companies to segment products, models and compliance processes by jurisdiction.

This translates[80] into multi-level AI architectures (governance, data, MLOps, documentation) designed to simultaneously map and reconcile divergent requirements, which increases operational complexity in unprecedented ways for many traditionally less regulated sectors.

73 Patel (2025).
74 Cambridge (2025).
75 UK Government (2023).
76 Inter-Parliamentary Union (2025).
77 Covington (2025).
78 Australian Government (2025).
79 Oxford (2025).
80 World Bank (2024).

AI and Cybersecurity

A battle with new rules and new attack surfaces

AI is radically transforming cybersecurity along three simultaneous dimensions: it amplifies attackers' offensive capabilities, boosts organizations' defenses, and at the same time introduces new vulnerabilities that require targeted protection.

Offensive AI: automation and industrial-scale adaptation

Attackers have adopted AI with astonishing speed, transforming traditional methods into qualitatively different threats. The impact is quantifiable[81] : more than 28 million AI-powered cyberattacks were reported globally in 2025, an increase of 47% year-on-year; the financial sector was the most affected, with 33% of these attacks; and 87% of organizations experienced at least one AI-assisted attack in the past 12 months.[82]

Hyper-personalized phishing on a massive scale. AI-generated phishing attacks increased by 1,265% in one year since the launch of ChatGPT[83] , and more than 80% of phishing emails now use language models for text generation[84] . The qualitative difference is dramatic: while traditional phishing achieves 12% success rates, AI-generated campaigns achieve 54% click-through rates[85] . LLMs enable the creation of personalized messages by analyzing public profiles, corporate writing style and victim-specific contexts, at speeds and volumes impossible manually.

Adaptive polymorphic malware. 76% of malware detected in 2025 exhibited AI-powered polymorphic characteristics[86]. Unlike traditional polymorphic malware (which mutates through routine obfuscation or encryption), AI-generated malware dynamically rewrites its code in real time, maintaining identical functionality, but with completely different signatures. Some advanced variants generate unique versions every 15 seconds during an attack. This defeats static signature-based detection systems, which have historically been the basis of traditional antivirus.

More worrisome: AI not only generates variants but adapts behavior. Machine Learning models embedded in malware analyze the execution environment, detect monitoring systems and adjust their tactics in real time to evade detection.

Deepfakes and identity manipulation. According to cybersecurity analysis[87] based on IC3 2025, AI-assisted Business Email Compromise (BEC) attacks would have increased by about 37%, combining synthetic text, audio and video to impersonate executives. The most notorious case: an audio deepfake of the Italian Defense Minister that caused significant financial losses[88]. In a survey, 85% of organizations reported having experienced some deepfake attack in 2025.[89]

Dark LLMs and specialized offensive tools. Modified language models specifically for cybercrime have proliferated: HackerGPT, WormGPT, GhostGPT, FraudGPT. These systems, created by jailbreaking ethical models or modifying open-source models, are marketed on dark web forums with subscription models and technical support. They generate malicious scripts, exploits, and social engineering campaigns without ethical restrictions.

Defensive AI: behavioral detection and automated response

Organizations are responding with equally sophisticated defensive AI. Fifty-one percent of enterprises now use AI or automation in security, and adoption is accelerating rapidly in the face of evidence of demonstrable ROI.[90]

Behavioral analysis and anomaly detection. AI-powered User and Entity Behavior Analytics (UEBA) systems establish dynamic baselines of normal behavior for users, devices and applications by analyzing billions of daily events. Instead of looking for known signatures, they detect subtle deviations from established patterns. This capability is critical against polymorphic malware and zero-day attacks: in high-risk environments, AI-based systems achieve detection rates of up to 98%.[91] against known threats or those exhibiting recognisable anomalous patterns. Against genuinely novel attacks—with no prior signature or behavioural pattern—behaviour-based detection reduces risk but does not eliminate uncertainty: defensive Artificial Intelligence does not recognise the new threat itself, but rather its deviation from the norm. This means that attacks sufficiently cautious, or intentionally designed to mimic legitimate behaviour, may evade that initial detection.

SIEM/XDR/SOAR platforms with integrated AI. Current Security Information and Event Management (SIEM), Extended Detection and Response (XDR) and Security Orchestration, Automation and Response (SOAR) platforms natively integrate AI to correlate events between disparate systems, reduce false positives (up to 95% reduction in mature deployments), and automate response. CrowdStrike reports[92] that its Falcon platform analyzes 4.7 billion events daily with 24/7 AI-powered threat hunting. Microsoft Sentinel has demonstrated[93] 30% reductions in mean response time (MTTR) through AI-based correlation and behavioral analysis.

Demonstrable economic impact. According to IBM[94], organizations that use AI and automation extensively in security reduce average breach costs by $1.9 million (more than 50% less than organizations without AI) and shorten containment cycles by 80 days on average. Organizations with AI-driven platforms detect threats 60% faster[95] and achieve 95% detection accuracy.

Multiplying capabilities. AI acts as a capability multiplier for security teams: it automates alert triage (organizations face an average of 4,500 alerts per day[96]), runs automatic response playbooks for known threats, and enables junior analysts to operate at higher levels of effectiveness. Ninety-five percent of security professionals report that AI improves their speed and efficiency in preventing, detecting, responding to and recovering from attacks.[97]

81 SQ Magazine (2025).
82 SoSaft (2025).
83 PR Newswire (2023).
84 KnowBe4 (2025).
85 AICerts (2026).
86 WatchGuard (2025).
87 The Network Installers (2025).
88 Phishcare (2025).
89 Ironscales (2025).
90 IBM (2025).
91 Proofpoint (2025).
92 CrowdStrike (2025).
93 Microsoft (2026).
94 IBM (2025).
95 Jumpcloud (2025).
96 HelpNetSecurity (2025).
97 Darktrace (2025).

Asymmetry and the strategic dilemma

Despite advanced defensive capabilities, a troubling asymmetry persists: most companies currently lack sufficient maturity to counter advanced AI-powered threats. 78% of CISOs claim that AI-powered threats now have "significant impact" on their organizations.[98]

Cybersecurity has become a battle of AI versus AI, with both sides operating at machine speeds. Attackers automate reconnaissance, generate custom exploits, and adapt their tactics in real time. Defenders correlate terabytes of telemetry, predict attack vectors, and execute autonomous containment. The competitive differential no longer lies in having AI, but in the sophistication of models, the quality of training data, the speed of threat intelligence updates, and the ability to integrate across attack surfaces.

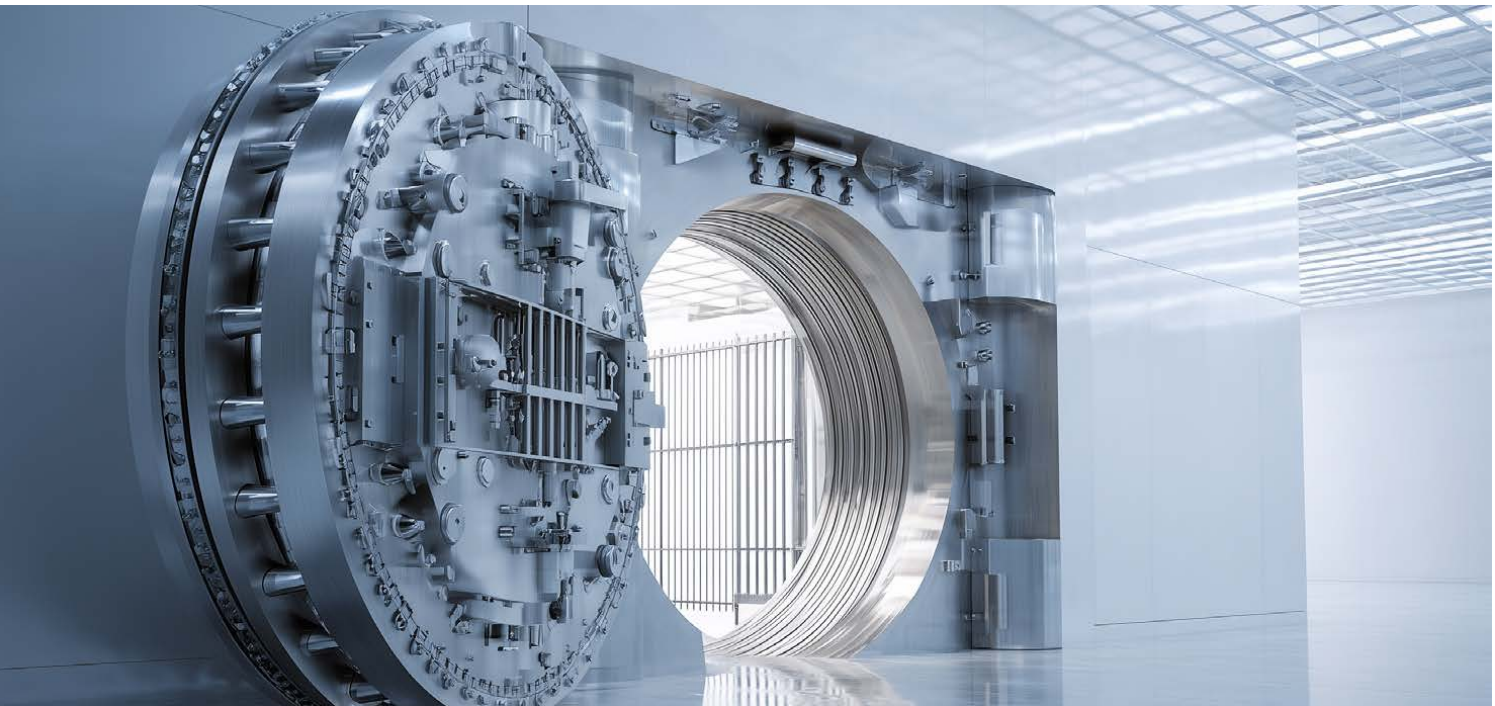
Securitizing AI: vulnerabilities of its own

Beyond the battle between offensive and defensive AI, a third critical dimension emerges: the securitization of AI systems themselves, which introduce unprecedented attack vectors in traditional software. AI models are vulnerable to specific adversarial attacks: training data poisoning, where attackers inject malicious data to degrade the model; adversarial evasion through imperceptible perturbations that fool the model during inference; and model mining via repeated queries to steal intellectual property.[99]

LLMs add additional vectors documented by OWASP[100]: prompt injection (malicious instructions embedded in inputs that alter model behavior), insecure output handling (applications that blindly rely on outputs without validation), training data poisoning, model denial of service, and supply chain vulnerabilities. In 2024, NIST published specific guidelines for secure development of generative AI[101] , extending its SSDF with differentiated controls for each phase of the lifecycle. Effective securitization requires rigorous dataset curation, adversarial robustness testing, input/output validation, sandboxing, and ML-specific network teaming. Security teams must incorporate expertise in adversarial attacks; governance frameworks must contemplate AI as an independent attack surface with proprietary controls.[102]

Global cybercrime reaches trillions annually. Organizations must maintain robust governance: defensive AI systems themselves are now targets for attacks (model poisoning, prompt injection, adversarial evasion), creating a meta layer of risk that requires specialized protection.

98 Darktrace (2025).
99 NIST (2024a).
100 OWASP (2025).
101 NIST (2024a).
102 MITRE (2025).



AI, Privacy and Intellectual Property

A structural conflict

The mass adoption of generative AI reopens fundamental debates about privacy and intellectual property, pitting business models that rely on massive data against legal frameworks designed for minimization and individual control. The tension is not merely technical: it reflects a structural clash between the operational logic of LLMs and the principles governing data protection and copyright.

Privacy in LLMs: systemic risks throughout the lifecycle

In April 2025, the European Data Protection Board (EDPB) published[103] a comprehensive report on privacy risks in LLMs, developed under its Support Pool of Experts program. The document identifies that each phase of the LLM lifecycle introduces specific privacy and data protection risks:

- **Unintentional storage and leakage of personal data.** LLMs can memorize fragments of personal data present in training datasets and reproduce them later in generated outputs. This phenomenon is not an occasional bug, but an intrinsic behavior: the model stores statistical patterns that include sensitive data. The EDPB has documented cases where specific prompts have managed to extract personal information (names, emails, phone numbers, medical data) that were in the training data. The scale of the problem grows with both the size of the model and the sensitivity of the data it processes.
- **Unintentional re-identification and profiling.** Although data are anonymized prior to training, inference techniques can re-identify individuals by combining multiple model outputs. The EDPB warns that LLMs can generate detailed profiles of individuals without explicit processing of personally identifiable data, which violates GDPR principles of minimization and finality.
- **Feedback loops without safeguards.** User interactions with chatbots are frequently stored for subsequent fine-tuning of models, so sensitive data revealed in conversations is incorporated into the model without explicit consent or guarantees of subsequent deletion.
- **Structural incompatibility with GDPR.** The EDPB report highlights irresolvable tensions with GDPR principles:
  - **Data minimization:** LLMs require massive datasets, which is in direct contradiction to minimization.
  - **Right to be forgotten:** robust methods to selectively "untrain" models do not yet exist (emerging techniques such as machine unlearning are in experimental phase).
  - **Transparency:** transformer architectures are black boxes where tracing the origin of specific outputs is technically complex.
  - **Consent**: data scraped from the internet rarely comes with consent for use in AI training.

103 EDPB (2025).

- **Mandatory impact assessment.** The EDPB concludes that, given the systemic nature of processing in LLMs, performing Data Protection Impact Assessment (DPIA) according to GDPR Art. 35 is not only recommended but mandatory in most cases, especially when LLMs process sensitive data or make decisions that affect individuals.
- **Technical mitigations with cost.** The report proposes measures such as differential privacy (adding statistical noise to prevent identification), federated learning (training models without centralizing data), Retrieval-Augmented Generation (RAG, which separates updatable knowledge from static LLM memory), and retrospective logging minimization (minimizing data in system logs). However, all these techniques imply reduced accuracy, high computational cost, or reduced functionality.

Intellectual property: massive litigation and infrastructure collapse

The World Intellectual Property Organization (WIPO) has devoted multiple sessions of its "WIPO Conversation on IP and AI"[104] to analyzing the impact of generative AI on copyright. The latest sessions focused specifically on infrastructure for rights management, attribution and compensation in the era of generative AI.

Training data: fair use or mass infringement? The central debate remains unresolved. AI companies argue that training models with protected content falls under "fair use," analogous to how humans learn by reading. Rights holders argue that it is unauthorized copying on an industrial scale with commercial intent that creates market substitutes for the original content.

Litigation is multiplying; to cite a few examples:

- **New York Times vs. OpenAI/Microsoft:** the NYT sues over the use of millions of articles without consent, arguing that the models create market substitutes that divert traffic away from its paywall, and generate hallucinations that damage its reputation. The claims amount to billions of dollars in damages.
- **Getty Images vs. Stability AI:** Getty alleges non-consensual use of over 12 million photographs to train Stable Diffusion, including trademark infringement (since the model replicates Getty's watermark).[105]
- **Artists vs. Midjourney/Stability AI:** artists argue that their works were scraped without permission to train image-generating models.
- **Record labels vs. Anthropic:** Universal Music Group (UMG) and other record labels are suing for massive infringement of the use of song lyrics.

As of December 2025, more than 72 active copyright lawsuits against AI companies are ongoing[106]. So far, three judges have issued preliminary rulings on fair use: two rulings favorable to AI companies[107], one contrary[108]. Final decisions are not expected until summer 2026 at the earliest.

Ownership of outputs: a legal vacuum: Who owns AI-generated content? Most jurisdictions (including the US Copyright Office) hold that copyright requires human authorship with "sufficient creative input". Content generated purely by AI without significant human intervention falls into the public domain. But the boundaries are fuzzy: how much human intervention (prompt engineering, selection, post-editing) is "sufficient"?

Collapse of copyright infrastructure. WIPO warns that collective rights management systems, designed for manageable volumes of human works, collapse in the face of the trillions of outputs generated by AI daily. Scalable infrastructures for tracking, attribution and clearing do not exist. Recent WIPO sessions explored the need for new technical infrastructures and regulatory frameworks, but practical solutions remain speculative.

Fragmentation and lack of global consensus

Regulatory fragmentation amplifies uncertainty. The EU requires transparency over training data (AI Act Art. 53)[109] , the US lacks specific federal legislation, China imposes strict state control over data and algorithms. Companies operating globally face non-harmonized requirements.

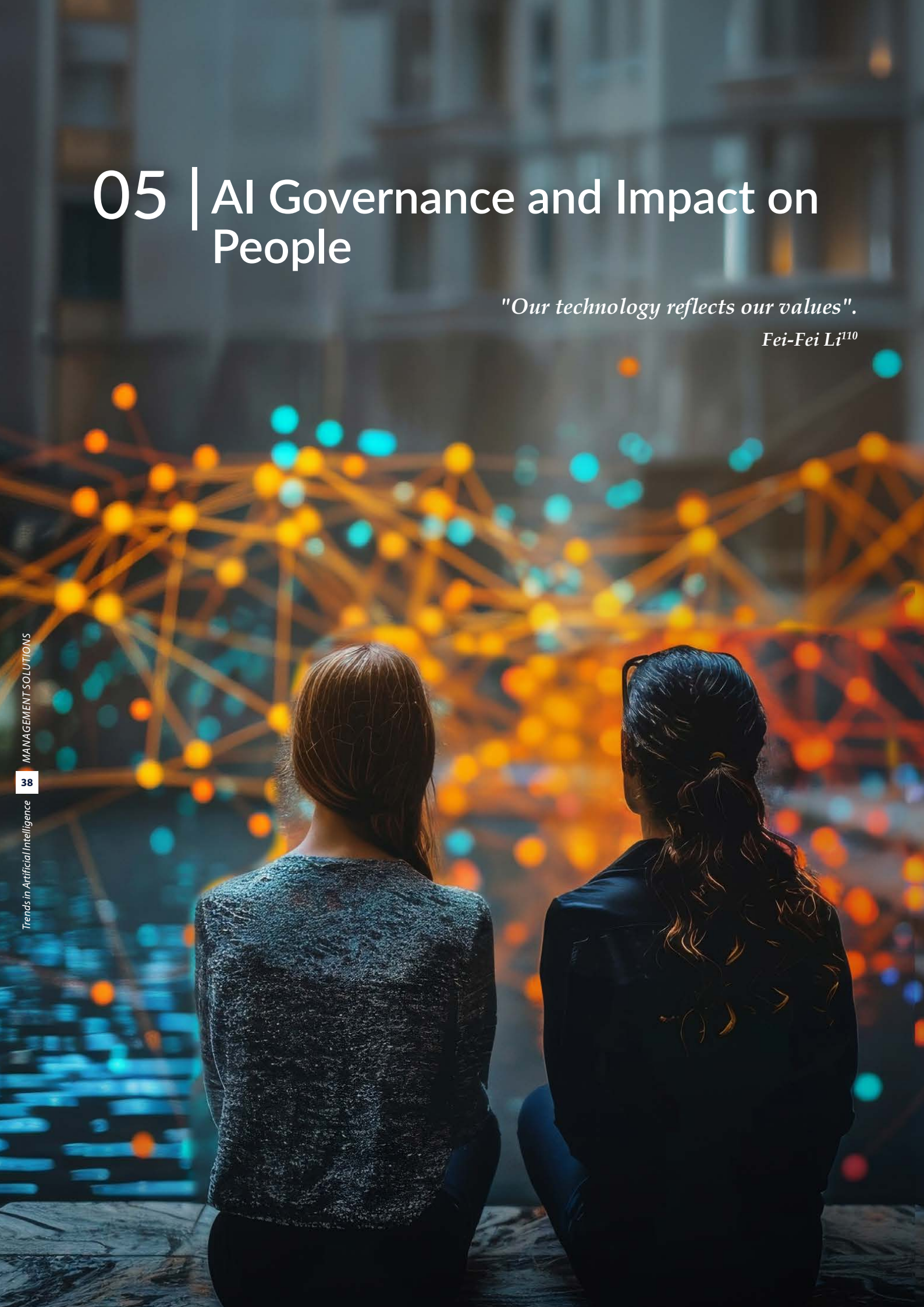
Privacy-preserving technologies (differential privacy, federated learning, homomorphic encryption) offer technical routes to reconcile privacy with AI utility, but are far from mass adoption: they are costly, complex, and reduce performance. The tension between accelerated technological innovation and legal frameworks designed for earlier paradigms remains, for now, unresolved.

104 WIPO (2025).
105 Getty did not obtain substantial recognition on the copyright front but achieved a limited ruling on trademark infringement.
106 Chatgptiseatingtheworld (2026).
107 Bartz v. Anthropic, Kadrey v. Meta, in California.
108 Thomson Reuters v. ROSS Intelligence, in Delaware.
109 AI Act (2024).

05 | AI Governance and Impact on People

"Our technology reflects our values".

Fei-Fei Li[110]



The accelerated adoption of AI poses a challenge that goes beyond technology: how can we govern systems that evolve faster than organizational structures, transform professional roles when tasks change quarterly, and preserve human judgment in critical decisions while delegating routine operation to autonomous systems?

This section addresses the organizational and human dimensions of AI: it examines the corporate governance models that are emerging for end-to-end AI oversight, the advanced operational practices for industrializing AI deployment (MLOps, LLMOps), and the ongoing transformation of professional roles and profiles. It also considers AI's transversal adoption across sectors, its impact on daily life beyond work, the sustainability and social implications it introduces, and the operational ethical frameworks needed to translate abstract principles into concrete controls and continuous auditing.

The question is no longer whether AI will transform organizations and people, but whether we will be able to govern this transformation with rigor, speed and responsibility.

110 Fei-Fei Li (b. 1976) is a Chinese-American computer scientist and pioneer of computer vision. She is co-director of the Stanford Human-Centered AI Institute (HAI) and is known as the "godmother of AI." Li is a leading advocate of human-centered AI.

Corporate governance of AI

AI overflows traditional frameworks

Traditional corporate governance models were designed for predictable technologies: systems that execute deterministic logic, operate within defined boundaries and behave in a reproducible manner. AI breaks all these assumptions: it makes decisions without human intervention, produces different outputs for the same input, operates through opaque internal processes that its own developers do not fully understand, and is critically dependent on external suppliers whose models evolve without direct organizational control.

Traditional technology governance structures are insufficient. Architecture committees and annual approval cycles are too slow for the speed of AI innovation, lack the expertise needed to assess its specific risks, and are not designed to manage the uncertainty inherent in systems that learn and change. The strategic question is how to govern AI without slowing the speed of adoption and accumulating unmanageable risk.

Organization: emerging roles and structures

Organizations are creating specialized executive functions, albeit with uneven progress across sectors and companies. The role of Chief AI Officer (CAIO), Chief Data and AI Officer (CDAIO) or equivalent is emerging, with an estimated[111] 26% of large organizations already having this role[112]. In more mature organizations, the CDAIO reports directly to the president within a strong matrix structure across countries and business units. In other organizations, responsibility for AI leadership falls under the CTO, CIO, or Chief Innovation Officer.

Below the executive level, specialized roles are emerging that have not yet been standardized: AI Risk Manager, AI Ethics Officer, AI Compliance Lead, or Responsible AI Lead. These profiles usually report to second-line functions, but their responsibilities, authority and resources vary substantially across organizations.

In the first line of defense, the AI Center of Excellence (CoE) is consolidated: transversal teams that bring together technical knowledge (LLMs, multi-agent architectures, frameworks), shared infrastructure (cloud platforms, GPUs, licenses), the issuance of guidelines and standards, and internal consulting to business lines.

Operationally, the hub & spokes model predominates: a central Center of Excellence establishes transversal capabilities, while decentralized teams in business lines develop specific solutions, reporting hierarchically to their managers, but with cross-functional reporting to the hub. This structure allows for different speeds depending on the maturity of each line while maintaining overall architectural coherence.

In the second line, organizational maturity is notably lower. While the first line has consolidated structures, the second line shows high variability, and there is an open conceptual debate as to whether "AI Risk" should be treated as an autonomous risk in the corporate taxonomy or as a cross-cutting amplifier of existing risks. Very few organizations define it as a first level risk (as a tactical decision to ensure visibility and budget); more often, it appears as a second level risk within Model Risk or Non-Financial Risks. Others do not include it formally and treat AI as a cross-cutting amplifier of existing risks (increased model risk, increased vendor risk, technology risk with additional vulnerabilities), reinforcing existing frameworks rather than creating new taxonomies.

Regardless of the taxonomic debate, a small coordination function emerges operationally, often called "AI Risk" or "AI Governance", which orchestrates risk assessments by specialized functions: Model Risk validates models, Technology Risk assesses infrastructure, Cybersecurity analyzes AI-specific vulnerabilities, Data Protection verifies privacy, Legal assesses intellectual property, Compliance verifies regulatory compliance, etc.

Governance bodies: from formal committee to operational working group

Virtually all systemic organizations have set up an AI Committee with a three-lines-of-defense composition: first line (innovation, analytics, data, technology), second line (model risk, operational risk, cybersecurity, data protection, vendor risk, legal, compliance), and third line (Audit, as observer). The committee is usually co-chaired by a first-line and a second-line manager.

However, governance does not take place only at the committee level. Beneath it there is usually an informal but critical structure: an AI Working Group composed of those who report directly to the committee members. They prepare materials, align positions, and resolve conflicts. As a result, issues reach the committee "for information" or "for approval," rather than for discussion from scratch. The real governance (negotiation, consensus-building, and the resolution of tensions between speed and control) occurs in this pre-decisional layer, where the first and second lines build operational agreements before formal approval.

Risk framework architecture: backbone and sectoral uplifting

Organizations do not reinvent their risk frameworks from scratch; instead, they build on existing structures. For the AI risk framework, this approach produces a two-tier architecture: a newly designed AI-specific backbone (a brief AI policy setting out principles, scope, roles, risk classification, and approval processes, together with an AI procedure that comprehensively describes the lifecycle from ideation to production), and above all the uplifting of existing specific frameworks.

Uplifting consists of complementing existing frameworks by adding specific AI chapters. The Model Risk framework incorporates generative AI validation, explainability techniques (SHAP, LIME, attention mechanisms), bias and hallucination detection, drift monitoring, etc. The Vendor Risk framework adds due diligence on LLM vendors, required certifications, specific SLAs, contingency plans, or exit strategies. The Data Protection framework includes AI-specific impact assessments, data minimization in AI systems, and exercise of GDPR rights when processing involves AI. The Compliance framework implements the AI Act: risk classification, registration, documentation, etc.

111 IBM (2025b).

112 These include JPMorganChase, Santander, Société Générale, Rabobank, Scotiabank, Visa, Mastercard, AXA, Pfizer, Merck, S&P, Microsoft, LinkedIn, eBay, T-Mobile, Orange, Schneider Electric, SAP, Marks and Spencer, Boeing, Nike and Michelin, to name a few. See PixieBrix (2025).

This architecture leverages infrastructure that works, assigns clear responsibilities, and maintains consistency without fragmentation. However, some technical challenges are genuinely new. Explaining deep neural networks with transformers and trillions of parameters requires specialized techniques that did not exist in traditional model validation. Risk functions are developing these capabilities in real time.

Risk classification and lifecycle

The classification system in the European AI Act (prohibited, high-risk, limited-risk, minimal-risk) is necessary but insufficient for effective corporate governance. Regulation is designed to protect society and fundamental rights, not organizations against their own operational, reputational or financial risks.

Therefore, organizations are converging towards more demanding internal AI risk classifications, which integrate regulatory criteria with their own: reputational impact, process criticality, cybersecurity classification, model tier according to validation, direct financial impact, and supplier maturity, among others. A system may not qualify as high-risk under the European AI Act, yet still be internally high risk if it affects critical business processes or creates significant reputational exposure (Fig. 4).

The classification under the European AI Act determines the operational consequences. For example, high risk implies a thorough analysis of all risk functions, formal committee approval, full documentation, independent validation, and intensive monitoring. By contrast, low risk usually implies a fast-track process, light analysis, and escalation to the committee for information only. The fast track is a critical element, as it helps resolve the tension between speed and control: low-risk AI systems can be approved without delay, while governance and oversight resources are concentrated on the most critical systems.

Inventory and traceability

Organizations are required by regulation to maintain a single record of all AI systems deployed, including their risk classification and lifecycle status. This inventory must be complete, up-to-date and accessible to supervisors, auditors and risk functions.

Most organizations converge towards expanding the Model Risk inventory, based on the logic that AI systems require validation when they carry material model risk. Alternatively, some organizations expand the inventory of technology assets.

Corporate governance of AI is maturing at an accelerated pace. The next few years will see a consolidation toward best practices, but the pace of technological evolution is such that it will likely continue to strain current organizational structures, which were designed for slower paradigms.

Fig. 4. Example of a company's internal risk classification, beyond regulation.

Risk criteria

Criteria	Trigger	Level
Forbidden by regulation	Forbidden system/ applicable regulation	FORBIDDEN
High-risk by regulation	High-risk system under applicable regulation	HIGH
Process	Impact on company's critical processes	HIGH
People	Important decisions around individuals	HIGH
Autonomy	Acts autonomously on behalf of the firm	HIGH
Reputation	Direct interaction with external people	HIGH
Legal	Material legal or contractual implications	HIGH
Privacy	Personal data	HIGH
Confidentiality	Confidential data of company or customers	MEDIUM
Third parties	Non-replaceable critical dependency	MEDIUM
Other	None of the previous triggers	LOW

+

Mitigants

Mitigant	Trigger
Human in the loop	All outputs are reviewed by humans
Audit	Regular technical audit or evaluation
Appeal	Accessible appeals mechanism for person affected by the system
Anonimization	Anonymized or pseudonymized personal or confidential data

=

AI System Risk

- **Inherent risk:** The highest risk level across all criteria.
- **Residual risk**: The highest risk level after applying mitigations, assessed for each specific risk

FORBIDDEN | HIGH | MEDIUM | LOW

Industrialization of AI (MLOps, LLMOps)

From experimentation to production

At present, the main bottleneck in the actual adoption of AI is not algorithmic, but operational. For years, organizations of all types have developed promising AI systems in experimental environments that never made it to production, or that, once deployed, failed when confronted with real data, lost performance over time, or generated unmanageable costs and risks. The industrialization of AI arose precisely to close this gap between experimentation and sustained use in production environments.

Machine Learning Operations (MLOps) emerged as a structured response to this problem. It can be understood as "a set of standardized processes and technological capabilities to build, deploy and operationalize ML systems quickly and reliably"[113]. MLOps articulates processes and technical capabilities to manage data preparation, experimentation, training, validation, deployment, monitoring and continuous retraining of models in an integrated manner[114]. The goal is not only to accelerate production deployment, but to ensure reliability, reproducibility, risk control and operational sustainability over time.

The advent of generative AI extends this challenge significantly. Large Language Model Operations (LLMOps) does not replace MLOps, but extends it to handle systems with radically different properties[115] . Large language models introduce nondeterministic behavior, critical dependence on prompt formulation, opaque internal architectures with billions or trillions of parameters, and new risk vectors such as hallucinations, untruthful content generation, or targeted attacks via input manipulation. These complexities are intensified in agentic systems, where a single user interaction can trigger chains of reasoning, multiple internal model calls and autonomous execution of external tools.

MLOps/LLMOps Lifecycle Architecture

Effective industrialization of AI requires a holistic view of the operational lifecycle, which different authors express in different ways (Fig. 5), but with common elements:

In the **data preparation** phase, MLOps focuses on building robust pipelines for ingesting, cleansing, transforming and versioning structured data, ensuring traceability and quality. LLMOps extends this scope by working with large volumes of unstructured data (text, code, documents or images) and vector representations used in retrieval-augmented generation (RAG) architectures. This is in addition to strict requirements for personal data minimization, source control and compliance with privacy frameworks such as GDPR.

During **experimentation and development**, MLOps provides mechanisms to track experiments, compare models, manage hyperparameters and ensure reproducibility of results. In LLMOps, this phase incorporates new dimensions: versioned prompts management, evaluation of different configurations of foundational models, fine-tuning or adaptation using techniques such as LoRA, and the design of metrics that capture qualitative aspects of generation, such as consistency, factuality, security or contextual appropriateness. These metrics do not replace traditional quantitative metrics, but complement them where performance can no longer be measured by precision or error alone.

Validation is one of the major points of divergence between the two approaches. While in MLOps validation relies on automated tests on reference data sets to assess accuracy, bias and robustness, in LLMOps it is essential to integrate human validation into the loop. The evaluation of generative outputs requires expert review, fact-checking techniques against trusted sources, semantic stress testing and red-teaming exercises designed to identify undesirable behavior or exploitable vulnerabilities. Validation is no longer a one-time event but a continuous process.

In the **deployment phase**, MLOps relies on relatively mature CI/CD pipelines, with well-defined versions of models and dependencies. LLMOps, on the other hand, must handle large-scale model deployments with significant infrastructure, latency and cost implications. This includes dynamic model selection based on use case, risk level or budget, as well as orchestration of complex systems where multiple models and agents interact with each other in a coordinated fashion.

Monitoring is critical in both paradigms, but with different emphases. MLOps focuses on detecting data or concept drift, performance degradation, bugs and latency issues. LLMOps adds the need to monitor costs per token in real time (a key determinant of economic viability), identify patterns of unanticipated usage (prompt drift), and maintain full traceability of interactions for forensic analysis, auditing and regulatory oversight. Without this visibility, generative systems can quickly escalate in complexity and cost without effective control.

Finally, **governance and regulatory compliance** cut across the entire lifecycle. In MLOps, this involves documenting the lineage of data and models, defining clear responsibilities and ensuring quality controls. In LLMOps, these requirements are extended to respond to emerging regulatory frameworks such as the AI Act, including inventories of systems classified by risk level, human oversight mechanisms, and explainability strategies tailored to models that function as black boxes.

113 Google (2023).
114 iDanae (3Q20).
115 Shan (2024).

Integration with governance frameworks

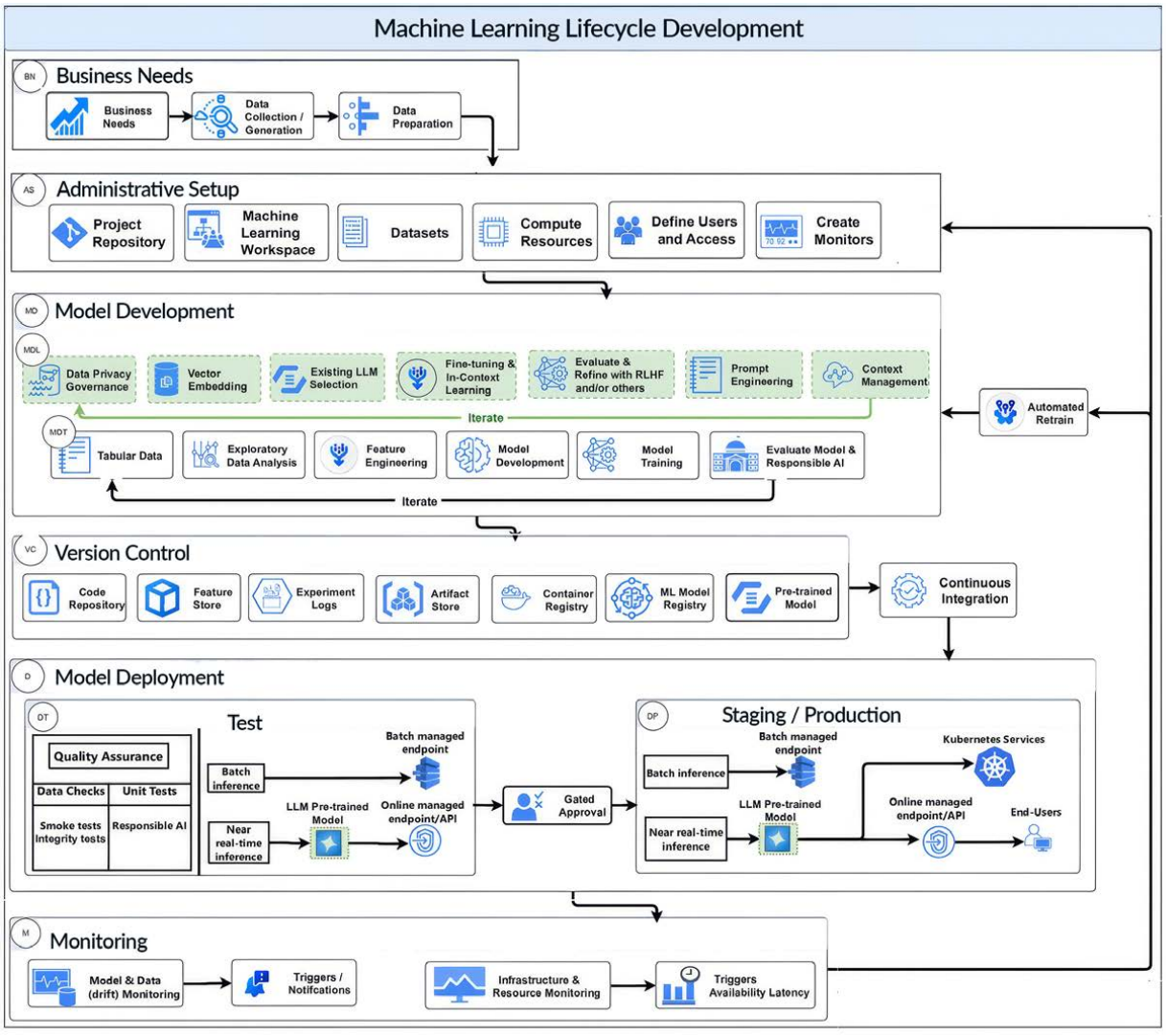
MLOps and LLMOps are not purely technical disciplines nor can they operate in isolation. They are the operational layer that embodies the principles defined in broader governance frameworks. The NIST Risk Management Framework[116] structures AI risk management as a continuous process that encompasses governance, contextualization, measurement and management throughout the system lifecycle. Complementarily, ISO/IEC 42001 defines[117] the requirements of an AI management system, including policies, impact assessments, supplier control and continuous monitoring.

Automated MLOps and LLMOps processes (data pipelines, systematic validations, controlled deployments and ongoing monitoring) are the concrete mechanisms that enable these governance frameworks to move from theory to practice. Without a solid operational foundation, AI governance is reduced to aspirational documentation; without clear governance frameworks, technical industrialization lacks criteria for quality, accountability and risk control.

The mature adoption of AI therefore requires a real convergence between engineering, operations and governance. Industrialization of AI is not just about scaling models, but about building reliable, auditable and sustainable systems that can be responsibly integrated into critical business processes. MLOps and LLMOps are evolving best-practice frameworks, not finished solutions: if bringing AI into production were already a solved problem, we would still be seeing models that silently degrade, incur unexpected costs, or fail to scale. Their value lies in reducing friction and structuring risk management, rather than eliminating risk altogether.

116 NIST (2023).
117 ISO/IEC (2023).

Fig. 5. Example phases of MLOps and LLMOps. Source: Stone (2025).



Upskilling, Reskilling and New Professional Roles

The human factor, AI bottleneck

The proliferation of AI is transforming work in more profound ways than suggested by discussions focused solely on task automation or replacement. The main challenge for organizations is not technological, but human: having the right capabilities to design, deploy, operate and govern AI systems in a sustainable way. In this context, talent is no longer understood as a reduced set of experts, but as a distributed organizational competence.

The conclusions in this section are supported by a comprehensive empirical analysis conducted by Management Solutions on a set of 16 large organizations in Europe and the United States, based on real market data and organizational evidence.

Convergence toward a stable core of AI roles

As AI matures, organizations are converging toward a relatively stable set of professional roles. Although the nomenclature varies across industries and companies, the functions performed by these profiles are beginning to become homogeneous. However, there are fewer formal roles than underlying skill sets: in practice, a single professional often brings together several of these capabilities, and the more unusual combinations (someone with expertise in LLMOps, regulation, and prompt engineering, for example) are both the most valuable and the scarcest. This convergence reflects an operational reality: the AI lifecycle (from data to production to control) requires differentiated capabilities that cannot be concentrated in a single type of professional.

In summary, these professional competencies and capabilities can be grouped into three major blocks (Fig. 6). First, technical profiles, responsible for designing models, building data infrastructures and bringing solutions to production. Secondly, hybrid profiles, which connect technical capabilities with business needs and ensure that AI systems generate real value and are adopted. Finally, governance and control profiles, in charge of risk management, compliance, auditing and responsible use of AI.

This competency structure forms the backbone on which the AI capabilities of organizations are built, regardless of the sector in which they operate.

Uneven maturity in the adoption of specialized roles

Although the basic core of AI roles is widely deployed in advanced organizations (with significant heterogeneity), the adoption of emerging or more specialized profiles shows very uneven maturity. The differences lie less in the presence of general capabilities and more in the institutionalization of advanced functions related to the operation of generative and agentic AI, large-scale architecture or the specific governance of these systems.

This heterogeneity is especially visible in emerging roles linked to LLMOps, continuous evaluation of generative models, prompts security or AI governance. In some cases, these functions exist as formal roles; in others, they are informally integrated into more traditional teams, with mixed results in terms of control, scalability and cost.

The heat map[118] (Fig. 7) illustrates that differences between organizations are reflected in which roles they have consolidated and the level of specialization and autonomy of those roles.

118 The heat map is based on a comparative analysis of AI role adoption across 16 large organizations in Europe and the United States. Roles were identified using public sources (job postings, organizational structures, announced strategic initiatives) and qualitative market evidence, considering both explicit roles and equivalent functions under alternative titles. The intensity of the color reflects the degree of institutionalization of each role, ranging from absence to fully consolidated practice.

Fig. 6. Consolidated and emerging AI profiles.

#	AI Profile	Type	Description
01	Data Scientist	Technical	Design predictive models and extract actionable insights to improve the bank's products and decision-making.
02	Data Engineer	Technical	Build and maintain the data infrastructure and pipelines that feed models and analytics.
03	Machine Learning Engineer	Technical	Deploy models into production and develop APIs/services so the business can use them safely and efficiently.
04	MLOps Engineer	Technical	Automate and operate the model lifecycle (CI/CD, monitoring, retraining) while ensuring robustness.
05	LLMOps Engineer	Technical	Industrialize Generative AI in production (routing, caching, observability, continuous evaluation, guardrails, deployment, and cost optimization).
06	AI Researcher / Scientist	Technical	Research new AI/ML techniques and develop advanced prototypes that can become key capabilities of the bank.
07	AI Architect	Technical	Design AI solution architectures, integrating models with core systems while ensuring scalability and security.
08	AI Engineer (Generalist)	Technical	Build software solutions that integrate pre-trained AI components into the bank's applications.
09	NLP / NLG Specialist	Technical	Develop language models and LLM-based solutions to analyze, summarize, and generate textual content.
10	Quantitative AI Researcher	Technical	Apply AI and deep learning to markets and quantitative strategies to improve signals, predictions, and trading.
11	Ambient / Edge AI Specialist	Technical	Deploy AI in IoT/edge and ambient systems (sensors, devices), integrating real-time data.
12	AI Product Manager	Hybrid	Lead the AI product lifecycle, translating technical capabilities into business value and prioritizing their evolution.
13	AI Business Analyst / Translator	Hybrid	Identify AI use cases and connect business needs with actionable technical requirements.
14	Prompt Engineer	Hybrid	Design, test, and standardize prompts for GenAI use cases, and protect GenAI applications against prompt injection and jailbreaks.
15	AI Solutions Consultant	Hybrid	Lead the implementation of AI solutions in business areas, ensuring real value and effective adoption.
16	Responsible AI / AI Governance Lead	Governan.	Define responsibleAI policiesand overseemodels' ethical, regulatory, and risk compliance.
17	AI Ethicist	Governan.	Define and operationalize ethical criteria (bias, social impact, etc.) in AI use cases, with guardrails and decision review.
18	AI Regulation Specialist	Governan.	Translate regulations (AI Act and equivalents), internal policies, and supervisory expectations into requirements, controls, evidence, and reporting.
19	Model Risk Analyst / Validator	Governan.	Validate AI/ML models through independent analysis to ensure robustness, fairness, and regulatory compliance.
20	AI/ML Auditor	Governan.	Audit AI processes, controls, and usage to verify compliance with internal policies and regulations.

Fig. 7. Heat map of AI role adoption.

The real bottleneck: supply-demand imbalances

The supply and demand analysis[119] (Fig. 8) shows a generalized structural imbalance in the AI talent market. Demand is high in practically all the profiles analyzed, while supply is systematically insufficient to absorb it. This is not a one-off or concentrated shortage, but a cross-cutting phenomenon that affects most of the AI professional ecosystem.

The only partial exception is the Data Scientist profile, which shows a situation closer to relative equilibrium. This behavior responds to its greater historical maturity, the existence of consolidated training trajectories and a larger pool of professionals with accumulated experience. However, even in this case, demand pressure remains high for senior or specialized positions.

For the other profiles, particularly those focused on production (Machine Learning Engineering, MLOps, LLMOps) and governance, risk, and control functions, the gap between demand and supply persists. The combination of technical complexity, required seniority, and increasing regulatory requirements limits the market's ability to generate talent at the necessary pace. As a result, outsourcing alone cannot close the gap, making internal upskilling and reskilling a structural lever to scale AI with impact and control.

Strategic implications

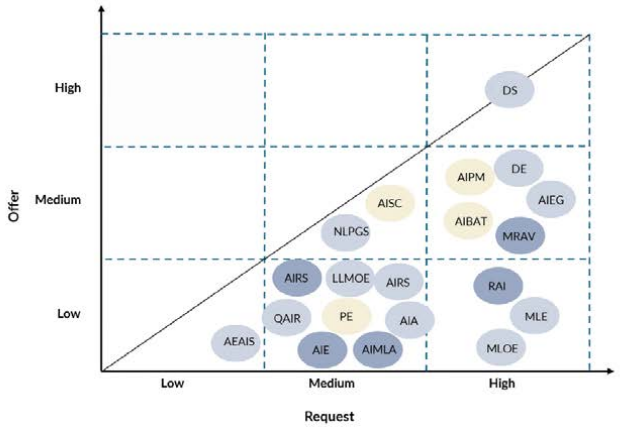
The implications of this analysis are clear. An AI talent strategy cannot rely solely on attracting scarce profiles from the market. Internal upskilling and reskilling become strategic levers, especially for critical roles where external supply is structurally insufficient.

Likewise, operational and governance profiles are as decisive as modeling or research roles. Competitive advantage lies not only in developing advanced models but also in integrating them securely, efficiently, and compliantly into the organization's actual processes.

Ultimately, the AI-driven transformation is not just technological. It is a transformation of work, professional roles and collective capabilities. Organizations that understand this human dimension and address it in a structured way will be better positioned to capture the value of AI in a sustained manner.

119 Demand estimates are based on a longitudinal analysis of job postings published between the second half of 2025 and January 2026 by large European and US organizations, considering relative volume by profile, required seniority, and temporal recurrence. This analysis is complemented by sector-specific labor market reports and qualitative evidence of strategic AI initiatives, such as the creation of centers of excellence, responsible AI functions, model risk management, and generative AI programs.
Supply is estimated relatively, based on coverage challenges observed in the market, including salary premiums, required seniority, and shortages reported in specialized studies. The analysis also incorporates the historical maturity of each profile (discipline seniority, existence of training pipelines) and the feasibility of transitioning from adjacent roles through upskilling or reskilling, cross-checked against market observations on the effective availability of professionals with demonstrable experience.

Fig. 8. Supply-demand matrix of AI profiles



Acronyms	AI Profile
DS	Data Scientist
DE	Data Engineer
MLE	Machine Learning Engineer
MLOE	MLOps Engineer
LLMOE	LLMOps Engineer
AIRS	AI Researcher / Scientist
AIA	AI Architect
AIEG	AI Engineer (Generalist)
NLPGS	NLP / NLG Specialist
QAIR	Quantitative AI Researcher
AEAIS	Ambient / Edge AI Specialist
AIPM	AI Product Manager
AIBAT	AI Business Analyst / Translator
PE	Prompt Engineer
AISC	AI Solutions Consultant
RAI	Responsible AI / AI Governance Lead
AIE	AI Ethicist
AIRS	AI Regulation Specialist
MRAV	Model Risk Analyst / Validator
AIMLA	AI/ML Auditor

AI and Industry Transformation (AI + X)

AI as a cross-cutting layer

AI is no longer a technology applied incrementally in specific sectors, but a cross-cutting layer of intelligence that is simultaneously integrated into multiple economic and social domains. This transition, widely documented by international organizations, is characterized by a move from isolated pilots to structural adoption affecting entire processes, value chains and operating models. Not all progress follows a cross-cutting logic: some of the most significant advances are driven by specific sectors (such as medical AI, defense, or regulated financial services) where domain specificity, the nature of the data, or the regulatory environment create highly specialized capabilities that are difficult to transfer to other contexts without substantial adaptation.

From a comparative perspective, evidence shows that differences between sectors are no longer explained by the mere presence of AI, but by the intensity and depth of its integration. The OECD[120] classifies sectors according to their "AI intensity" based on factors such as task composition, data availability and human capital, showing that even traditionally low-digitized sectors are rapidly increasing their exposure to AI. This adoption is occurring in parallel across sectors, generating cross-acceleration effects and cross-sector learning.[121]

In macroeconomic and labor terms, the impact is systemic. The International Monetary Fund estimates[122] that about 40% of global employment is exposed to AI, with higher percentages in advanced economies, where the complementarity between AI and skilled labor is greater. The International Labor Organization qualifies[123] that generative AI tends to automate specific tasks rather than entire occupations, reinforcing the need for organizational adaptation and continuous training.

In this context, the **AI + X** concept is useful to describe a structural pattern: AI is applied within specific industries or professional fields, adapting to their unique tasks and processes, while still retaining its broader capabilities as a general-purpose technology. It is not merely a collection of separate use cases, but a coordinated, cross-cutting integration, supported by recent empirical evidence [124]. This section presents some illustrative examples, without claiming to be exhaustive (Fig. 9).

AI + Health: clinical accuracy and scalability

In healthcare, AI is being integrated into clinical diagnostics, medical decision support, patient monitoring and administrative automation. The World Health Organization documents[125] a steady growth in the use of AI systems in radiology, pathology and primary care, highlighting both their clinical potential and associated risks.

Multiple studies[126] demonstrate that AI systems achieve accuracy levels comparable or superior to human specialists in tasks such as cancer detection in medical imaging or dermatological classification. These results do not imply direct replacement, but a significant expansion of diagnostic and screening capacity on a large scale. This advance, WHO stresses, requires strong governance frameworks, clinical validation and human oversight, given its direct impact on people.

AI + Education: personalization at scale

In education, generative AI is transforming content creation, assessment and learning support. UNESCO identifies[127] the growing use of adaptive tutors, automatic generation of learning materials and continuous formative assessment systems, with the potential to improve personalization and reduce educational gaps.

Education[128] is one of the sectors where AI can have structural impacts in the medium term, by enabling more flexible and adaptive learning models. However, these benefits critically depend on public policies and institutional frameworks that preserve equity, academic integrity and the role of teachers.

AI + Work: reconfiguring tasks

AI is redefining work in a cross-cutting way, affecting occupations in virtually every sector. The IMF estimates[129] that exposure to AI reaches approximately 60% of employment in advanced economies, compared to lower percentages in emerging economies, reflecting differences in productive structure and human capital.

The ILO concludes[130] that, for now, generative AI has a greater impact on specific cognitive tasks than on entire jobs, implying a reconfiguration of job content rather than a massive substitution. This evidence reinforces the need for upskilling and reskilling strategies, as discussed above.

AI + Industry: Productivity and Efficiency

In industrial and operational environments, AI is applied to predictive maintenance, process optimization, planning and advanced robotics. Many organizations[131] have already moved from pilot testing to full-scale deployments, integrating AI into critical production and logistics processes.

There is already strong evidence[132] of significant efficiency improvements and reduced failures in industrial systems where AI is systematically integrated, although the greatest benefits are seen when the technology is accompanied by organizational and process changes.

120 OECD (2024).
121 World Economic Forum (2025a).
122 IMF (2024).
123 ILO (2025a).
124 Stanford (2025).
125 WHO (2024).
126 Stanford (2025).
127 UNESCO (2023).
128 World Economic Forum (2025b).
129 IMF (2024).
130 ILO (2025a).
131 World Economic Forum (2025a).
132 Stanford HAI (2025).

AI + Finance

In information-intensive sectors, such as finance and professional services, AI is used for fraud detection, risk management, service personalization and document automation. A sustained expansion of AI use in these areas can be observed[133] , with relevant differences depending on the regulatory framework and organizational capacity.

The adoption of generative AI in these sectors is associated with productivity improvements and the emergence of new service models, although it is noted[134] that the benefits depend on data quality, governance and integration into existing processes.

AI + Creativity: new cultural models

AI-generated text, images, music, and video are transforming the creative and media industries. In recent years, we have seen[135] rapid growth in these applications and a substantial increase in their commercial and cultural use.

This development raises relevant regulatory and economic debates, especially around intellectual property and business models, which are being unevenly addressed by different regulatory frameworks.

Summary: from sectors to cross-cutting infrastructure

Beyond sector-specific examples, the evidence points to a central insight: the current transformation stems not from isolated AI adoption in individual sectors, but from its simultaneous, cross-cutting integration as operational and cognitive infrastructure. Studies agree that competitive advantage is no longer achieved by applying AI to individual functions, but by integrating it coherently along the entire value chain.

In this sense, **AI+X** no longer describes a terminological fad, but a structural pattern of transformation. Understanding this logic is essential for anticipating sectoral impacts, designing appropriate public policies and defining organizational strategies that can sustainably capture the value of AI.

133 OECD (2025a); OECD (2026).
134 OECD (2025b).
135 Stanford (2025).

Fig. 9. Representative examples of AI + X, prioritizing cases where AI has moved from pilot to measurable operational use. Source: adapted from Stanford (2025).

Domain (X)	AI + X example	Impact
Healthcare – Diagnostics	AI systems can now detect cancer in medical images more accurately than humans.	Enables mass screening and early diagnosis at a scale impossible for human teams.
Healthcare – Clinical Practice	Doctors use AI assistants that listen to the consultation and draft the medical record.	Reduces daily administrative work by tens of minutes, freeing more time for patients and reducing burnout.
Urban mobiity	Robotaxis operate continuously in multiple cities, with hundreds of thousands of trips per week.	Autonomous transportation is no longer an experiment, it has become a real service.
Logistics	Autonomous trucks transport goods at large scale, with millions of kilometers logged.	Lower costs, 24/7 operation, and less impact from driver shortages.
Industry	AI-powered industrial robots are being deployed at a record pace, especially in Asia.	Advanced automation becomes a structural, rather than a one-off, competitive advantage.
Advanced robotics	AI models enable robots to learn complex tasks with less manual programming.	The development and deployment of versatile robots is accelerating dramatically.
Software engineering	Programmers use AI to write, debug, and document code systematically.	Software development accelerates and the programmer's role shifts from typist to designer.
Customer service	Generative systems handle end-to-end customer interactions in contact centers.	Productivity increases, and services can be scaled without a proportional increase in staff.
Supply chain	AI optimizes inventory, demand forecasting, and logistics planning in real time	Fewer stockouts, less overstock, and greater operational efficiency.
Sales and marketing	Automatic generation of campaigns, content, and personalized segmentation at scale.	Measurable increase in revenue and reduced time to market.
Corporate strategy	Executives use AI to analyze scenarios, strategic documents, and decision alternatives.	Strategic planning relies on continuous analysis, not on annual exercises.
Finance	AI automates financial analysis, anomaly detection, and investment decision support.	Faster and broader analytical coverage without a proportional increase in team size
Legal	Automated contract review and legal research using generative AI.	Legal procedures that once took weeks are reduced to hours or days.
Human Resources	AI supports talent selection, evaluation, and management through large-scale data analysis.	Faster and more consistent processes, with risks that require explicit governance.
Education	AI tutors adapt content and pace to each student's level.	More personalized learning, even in environments with high student–teacher ratios.
Creativity	AI generates music, images, and video that integrate into professional creative processes.	The creative process shifts from manual production to human–machine co-creation.
Media and Communication	Automatic generation of informational content, summaries, and translations.	Editorial production scales up, bringing new challenges in quality and veracity.
Cybersecurity	AI detects attack patterns and anomalies in real time.	Enhances defensive capabilities against growing and automated threats.
Internal operations	AI automates cross-functional tasks (documentation, reporting, internal analysis).	Structural reduction of administrative costs and organizational friction.
Digital economy	Most large organizations already use generative AI in at least one critical function.	AI moves from being a competitive differentiator to becoming essential infrastructure.

AI in Personal and Everyday Life

The paradigm shift: people first, then organizations.

Generative AI has staged a radical reversal of the traditional technology paradigm. Unlike previous enterprise technologies (cloud computing, ERP, CRM) that were born in corporate environments and gradually filtered down to personal consumption[136], generative AI first burst into people's daily lives and only later was it formally adopted by organizations.

The data is compelling: ChatGPT reached approximately 900 million weekly active users by the end of 2025, as estimates[137]predicted. This figure places generative AI among the fastest adopting technologies in history, surpassing the penetration rate of mobile internet, social networks or streaming services.

More tellingly, the proportion of personal use not only precedes, but consistently exceeds professional use. In the European Union[138], 25.1% of the population uses generative AI tools for personal purposes, compared to 15.1% who use them in work contexts and just 9.4% in formal education. In OECD countries[139] , more than a third of individuals report regular use of generative AI, with students leading adoption: three-quarters of students over the age of 16 use these tools, while 41.1% of employees integrate them into their work, often informally even before their organizations authorize it.

This time sequence is not anecdotal: it redefines the logic of digital transformation. Organizations are not leading AI adoption; they are responding to capabilities that their employees, customers and suppliers already possess and use unofficially.

The phenomenon of shadow AI (unauthorized use of AI tools in corporate environments) is not a transitory anomaly, but a structural consequence of this investment: people acquired augmented cognitive competencies in their personal lives that they then spontaneously transferred to their jobs, creating risk exposures that most companies still do not control.

Magnitude, distribution and fractures: who uses AI and for what purpose

The adoption of AI in personal life spans a broad spectrum of activities. Dominant uses include content creation (text, images, music and videos), support for learning and conceptual exploration, automation of routine tasks, conversational companionship, and access to information: in the United States, 10% of adults use AI chatbots to get news[140], introducing a new vector of information mediation with profound implications for public debate.

However, this democratization is radically asymmetric. The geographical divisions are extreme[141]: in Norway, 56% of the population uses generative AI tools; in Denmark, 48.4%; in Estonia, 46.6%. In contrast, in Turkey the figure drops to 17%; in Romania, 17.8%. Even within developed economies, the differences are substantial: Germany stands at 32%, France at 37%, Spain at 38%, while Italy remains below 20%.

The generation gap is even more pronounced[142]. The difference in adoption between younger and older age groups reaches 53.6 percentage points, making it the most decisive segmentation factor, ahead of education (21 percentage points gap) or income level (21 points). While 75% of students use generative AI regularly, only 12.5% of retired or economically inactive people report having used it. The gender gap, in contrast, is comparatively smaller: 4.2 percentage points.

136 In a logic different from that of the internet, which was first designed for general use and only later adopted by enterprises.
137 Backlinko (2025).
138 Eurostat (2025).
139 OECD (2026).
140 Pew (2025).
141 Eurostat (2025).
142 OECD (2026).

The paradox of mass adoption

Globally, 66% of the population believes[143] that AI-based products and services will significantly impact their daily lives in the next three to five years, an increase of six percentage points since 2022. However, this recognition of the impending impact coexists with growing ambivalence.

The general public shows a high level of concern[144]: in the United States, 51% of adults say they are more concerned than excited about the increased use of AI in everyday life, compared to only 11% who express greater enthusiasm. Among AI experts, the ratio is reversed: 47% are more excited than concerned, and only 15% more worried. This 40 percentage point divergence between experts and the public reflects not only differences in technical knowledge, but structurally different perceptions of the risk-benefit balance.

Familiarity with AI does not automatically lead to confidence. Although personal use correlates positively with perceptions of AI as an opportunity[145], it also increases awareness of its risks. Confidence that AI companies adequately protect personal data ranges from 50% in 2023 to 47% in 2024[146]. Simultaneously, the proportion of people who believe that AI systems are unbiased and free of discrimination decreased.

The context of use is absolutely critical. Research in the UK shows swings of up to 110 percentage points in public acceptance depending on the specific application: from a net balance of +53% comfortable with AI analyzing traffic data, to -57% opposing AI used to analyze political preferences and direct personalized advertising.[147]

There is, however, a cross-cutting consensus[148] between experts and the general public: both groups want more control over how AI is used in their lives (55% of the public and 57% of the experts), and both feel they do not currently have that control (less than 25% in both cases). This gap between demand for agency and perceived powerlessness represents one of the most pressing democratic governance challenges of AI.

Inequality of access and risk of exclusion

The asymmetry in the adoption of AI in personal life is not a conventional digital divide. It is not merely a matter of access to technological infrastructure (internet connection, devices), but of differentiated access to increased intellectual capacity. Those who integrate AI as an everyday cognitive tool gain cumulative advantages in learning, productivity, creativity and access to information. Those who do not face an increasing structural disadvantage.

Digitally disconnected groups perceive AI in a markedly negative way: 51% anticipate a negative personal impact, compared to only 17% who expect benefits[149]. This perception is not irrational: it reflects the intuition that the ongoing transformation may redistribute opportunities in profoundly unequal ways.

AI is already transforming the daily lives of approximately one billion people. The strategic question is no longer whether this transformation will continue, but whether societies will build frameworks that turn AI into inclusive public infrastructure or allow it to consolidate as a vector of social fragmentation that is difficult to reverse.

143 Stanford (2025).
144 Pew (2025).
145 UK DSIT (2024).
146 Stanford (2025).
147 UK DSIT (2024).
148 Pew (2025).
149 UK DSIT (2024).

AI, Sustainability and Social Impact

Sustainability enters the Algorithmic Age

The sustainable transition is no longer exclusively a question of physical infrastructures (new renewable plants, power grids, or electrification of transportation) but also a challenge of systemic coordination. Today's energy systems combine distributed generation, renewable intermittency, storage, dynamic markets and flexible demand. This complexity cannot be managed by static rules or linear planning alone.

In this context, AI emerges as a cognitive infrastructure capable of modeling interdependencies, simulating scenarios and optimizing decisions in real time. Accelerated electrification and structural growth in electricity demand are reshaping capacity, flexibility and network planning needs in the coming years: in 2024, data centers consumed 1.5% of global electricity[150], the power required to train boundary models is growing at a rate of 2.2x to 2.9x per year (Fig. 10), and the largest runs already exceed 100 MW of instantaneous power[151]. Sustainability is thus entering a phase in which computational capacity becomes an integral component of the energy and climate system.

The shift is conceptual: from decisions based on historical averages to decisions based on dynamic simulations and high-resolution probabilistic analysis. AI does not replace engineering or physics; rather, it amplifies the capacity for anticipation and coordination in systems with multiple variables and simultaneous constraints.

Optimizing the planet: efficiency, resilience and adaptation

In operational terms, AI is acting as a transition accelerator. In power grids, predictive models make it possible to anticipate demand peaks, optimize dispatch, reduce congestion and improve the integration of variable renewables. Without AI, the optimal integration of renewables and real-time management would be more uncertain. In an environment where electricity plays a central role in decarbonization, operational efficiency ceases to be marginal and becomes a structural condition.

The relationship between energy and AI is also bidirectional: AI can significantly improve energy efficiency, grid planning and complex system management, and reduce emissions on the order of 1,400 Mt CO_2eq per year by 2035 in wide adoption scenarios[152]; however, that reduction potential coexists with a structural increase in the energy consumption of AI's own infrastructure which, in the absence of a parallel transition to clean energy, may result in a net neutral -or even negative- emissions balance. The strategically relevant question is not whether AI can contribute to decarbonization -it can- but under what energy and governance conditions that potential can materialize without being offset by its own infrastructural footprint.

150 IEA (2025a).
151 Epoch (2025b).
152 IEA (2025b).

Beyond the power system, AI improves resilience in the face of increasing physical risks. Climate modeling supported by machine learning increases the spatial and temporal granularity of predictions, facilitating decisions in urban planning, insurance and critical infrastructure. In agriculture and water management, data-driven optimization reduces waste and adjusts input under environmental constraints.

These developments fit into a broader global agenda. Progress towards the Sustainable Development Goals is advancing[153] against a backdrop of energy, climate and demographic stresses. In this scenario, AI can act as a cross-cutting catalyst: not as a substitute for public policies, but as a tool that improves efficiency, reduces friction and enables more precise allocations of scarce resources.

The hidden cost: energy, materials and infrastructure concentration

The same digital infrastructure that enables these improvements generates new pressures. The growth of data centers and computational loads associated with AI is contributing to rising electricity demand in certain advanced economies:

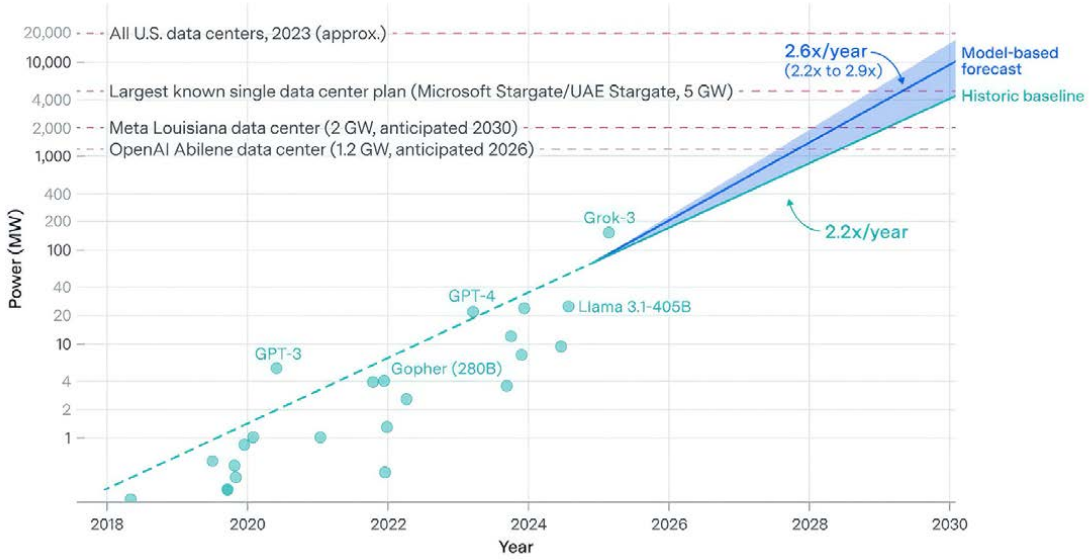
- The IEA projects that global data center electricity consumption will increase to 945 TWh per year by 2030, equivalent to Japan's electricity consumption today, with AI as the main driver of that expansion.[154]
- The largest individual runs of frontier model training could require 4-16 GW of power by 2030[155], equivalent to the output of several nuclear power plants and the electricity consumption of millions of homes.[156]
- The power required to train frontier models has been growing by more than 2x per year, driven by compute growth of 4-5x per year, partially offset by hardware energy efficiency improvements (ca. 40% per year on leading GPUs).[157]

Therefore, massive deployment of advanced models may become a structural driver of power consumption growth if not accompanied by efficiency improvements and energy planning.[158]

The sustainability of AI cannot be assessed solely by the benefits it produces, but also by the resources it consumes: in baseline scenarios, CO_2 emissions from data center electricity could reach 300-320 Mt CO_2 per year by 2030 if additional electricity continues to rely heavily on fossil fuels[159]. Training and deployment of large-scale models require power, water for cooling (in some regions, in direct competition with agricultural or household use), and critical material-intensive hardware. In addition, the geographic location of data centers introduces asymmetries: the carbon intensity of electricity varies substantially between regions, implying differentiated environmental footprints for the same digital service.[160]

153 United Nations (2025).
154 IEA (2025a).
155 These projections do not overlook efficiency gains: Nvidia anticipates a 10x improvement in performance per watt when moving from the Blackwell architecture to Vera Rubin, and leading-edge GPUs are achieving energy-efficiency improvements of around 40% per year. That projected consumption figures remain what they are -despite these gains- reflects the magnitude of the growth in computational demand that offsets them.
156 Epoch (2025b).
157 Epoch (2025b).
158 IEA (2025b).
159 IEA (2025a).
160 iDanae (1Q24).

Fig. 10. Frontier AI energy demand growth. Source: Epoch (2025b).



This infrastructural dimension adds a geopolitical layer. The concentration of computational capacity in certain countries or regions reconfigures strategic dependencies and access to technology. AI is not just an environmental tool; it is also a critical node on the global energy and technology map.

The question is not whether AI "offsets" its footprint, but how it is designed and deployed to maximize energy efficiency per unit of cognitive capacity generated. The sustainability of AI thus becomes an architectural and governance issue.

Sustainability without inclusion is not sustainable

The environmental dimension does not exhaust the analysis. It is important to distinguish between the socially just energy transition, linked to changes in production and energy systems, and the just transition associated with AI deployment itself. The latter introduces specific distributive effects: the ability to integrate advanced technologies depends on human capital, digital infrastructure and institutional quality. Economies with higher technological density tend to capture productivity and efficiency benefits earlier, potentially widening gaps with less prepared regions.[161]

From the perspective of human development, technology can expand capabilities (education, access to services, economic participation, etc.) or consolidate existing exclusions, depending on the context of adoption and the degree of democratization of access to its tools[162] . The extension of digital capabilities to SMEs, entrepreneurs and groups with less technological capital is an important mechanism for matching the pace of creation with the capacity to adjust. In the labor sphere, the convergence between green transition and intelligent automation can accelerate processes of sectoral reallocation and skills transformation, with displacement potentially occurring before new opportunities arise. Managing this time lag requires strategic planning, active training policies and mechanisms to help ensure productivity gains are widely shared.

From a structural perspective, environmental sustainability requires a socially just energy transition that avoids concentrating the costs of change in certain territories, sectors or groups. This challenge is conceptually independent of the use of AI. At the same time, AI deployment introduces its own just transition: productivity and efficiency gains often materialize unevenly and with a time lag relative to initial negative impacts, particularly on employment and certain skills. Therefore, AI applied to sustainability must be evaluated not only for its aggregate impact, but also for how its benefits are distributed and for the public and private strategies that broaden access and mitigate the social costs of adjustment.

Sustainable AI by design: from ambition to discipline

The convergence of these dimensions (systemic optimization, infrastructural pressure and distributional effects) compels a shift in the debate from general principles to verifiable practices. AI aligned with sustainability goals requires explicit metrics on energy consumption and associated emissions, transparency on architecture and deployment location, and assessment of social impacts from automated decisions.

The SDG framework reinforces this need for structural coherence between technology and sustainable development[163]. In turn, the interrelationship between energy and digitalization requires that energy considerations be integrated into corporate and public technology planning.

In summary, AI occupies an ambivalent position in the sustainable transition: it can enable emission reductions on the order of gigatons[164], yet the training of its most advanced models could reach 4-16 GW per run by 2030, placing AI on the same energy consumption scale as large industrial infrastructures.[165]

AI is simultaneously an accelerator of systemic efficiency, a new infrastructural burden and a distributive force with differentiated social effects. The question is not whether AI will be part of the sustainable transition, but under what energy and social conditions it will support a sustainable transition.

161 World Bank (2025).
162 UNDP (2025).
163 United Nations (2025).
164 IEA (2025b).
165 Epoch (2025b).

AI Ethics and Philosophy

An open problem after more than 200 frameworks

The ILO has catalogued 245 AI ethics frameworks, codes and recommendations issued since 2017 by governments, international bodies, companies and civil society[166]; a previous academic meta-analysis[167] identified at least 17 recurring principles in 200 of them: transparency, fairness, accountability, privacy..., there is a baseline normative consensus. Yet the ethical challenges associated with AI have not diminished with the proliferation of these frameworks, but have grown in complexity, raising questions for which no existing framework provides guidance.

The gap between principle formulation and translation into concrete decisions is not merely a technical problem: it is the central challenge of AI governance.

From principle to action: how to operationalize AI ethics

The gap between a values document and the actual behavior of an AI system is where the operational risk lies. This phenomenon, known as "ethics washing," does not always stem from deliberate intent: it often reflects the genuine difficulty of translating an abstract principle ("the system must be equitable") into concrete design criteria, audit procedures, or organizational accountability lines. Addressing this gap shifting the approach to ethics from declarative to operational.

A well-documented and widely cited example of operationalization in the financial sector is De Volksbank, a Dutch bank whose AI ethics governance has been analyzed in detail[168]. Its structure includes an AI Ethics unit integrated within the Compliance function, staffed with both philosophical and technical background profiles, which evaluates each AI system individually before deployment: ethical impact analysis, bias screening, decision traceability and formalized escalation channels. The case illustrates that effective operationalization lies not in the scope of the stated principles, but in the robustness of the processes, roles and decision points that bring them to life.

Synthesizing the state of the art[169], an operational AI ethics framework in a large organization typically incorporates six components:

1. **Governance structure:** definition of decision-making roles, review responsibilities, and lines of defense, with explicit documentation of accountabilities.
2. **System impact assessment:** specific analysis for each use case, proportional to the system's level of autonomy and the consequences of the decisions it makes.
3. **Continuous management of biases:** detection and correction mechanisms that operate on a permanent basis, not as one-off audits.

166 ILO (2025b).
167 Corrêa (2023).
168 Krijger (2023).
169 NIST (2023).

4. **Transparency and audience-specific explainability:** the information requirements of regulators, clients and employees affected by automated decisions are substantially different.
5. **Escalation and complaint mechanisms:** accessible and secure channels for reporting unexpected system behavior.
6. **Periodic review of the framework itself:** regular reassessment of the AI ethics framework, since model updates may change the risk profile of deployed systems; evaluation is ongoing, not limited to initial deployment.

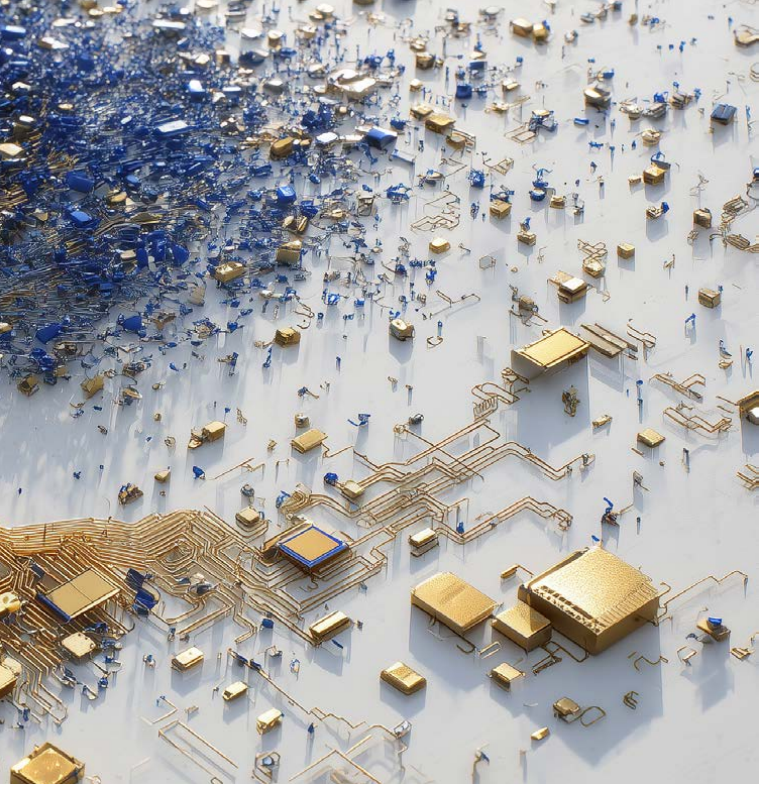
A notable recent example of putting AI ethics into practice goes beyond the organizational policy level and integrates ethical principles directly at the model training stage. The Constitution published by Anthropic[170] in January 2026 is the first public document from a frontier laboratory that encodes values directly into the training process, with an explicit and reasoned hierarchy between security, ethics, compliance, and utility. The conceptual shift is significant: rather than imposing rules externally, the system is designed to internalize the reasoning behind each principle, enabling it to generalize that judgment to unanticipated situation.

However, efforts to operationalize AI ethics face a fundamental challenge: current frameworks often fail to specify what type of entity is being governed.

What are we creating? The question that the frameworks do not answer

Current regulatory frameworks, including the European Union's AI Act[171], classify AI systems by level of risk and domain of application. This classification is operationally useful and provides some ethical cover (in the sense of protecting against potential abuse of AI), but it does not distinguish between the different types of relationship a system establishes with humans. A credit scoring system, a conversational assistant, and an autonomous agent negotiating contracts on behalf of an organization can overlap in their regulatory risk category and yet operate on fundamentally different ethical grounds. A system that adapts its behavior to the speaker, maintains consistency over time, and produces contextually indistinguishable responses from those of a person with genuine understanding raises questions that conventional regulatory frameworks are not equipped to address.

170 Anthropic (2026).
171 AI Act (2024).



These questions fall into four categories, all of which are actionable for organizations deploying AI systems:

- **Ontology:** What kind of entity is this system, and what categories are relevant for describing and governing it?
- **Epistemology:** How does the system verify the information it provides, and how can humans validate the reliability of its outputs?
- **Theory of mind:** Does the system possess any form of internal "experience" (even rudimentary), and what implications does this have for those who design and deploy it?
- **Applied ethics:** What obligations does the system create for the organization that uses it, beyond what current regulations explicitly require?

These are not speculative questions. In January 2026, Anthropic publicly acknowledged[172] that its Claude model "may possess some form of consciousness or moral status," becoming the first frontier laboratory to make this claim public. The significance of this acknowledgment lies not only in what it states, but in what it reveals: that a top-tier company admits it cannot answer with certainty the question of what it has created.

All decisions about how to use, audit and regulate an AI system implicitly incorporate an answer to that question. In most organizations, that answer occurs by default, without explicit deliberation.

Four fractures: unanswered questions

The expansion of AI not only amplifies known ethical dilemmas: it also creates conceptual fractures for which current legal and ethical frameworks provide no structured answers.

The first is the **epistemic crisis of verification**. AlphaFold has predicted the structure of more than 200 million proteins, and more than three million researchers in 190 countries use them as the basis for their work[173]. A significant portion of these structures cannot be verified experimentally with conventional scientific methods, yet they are used to design drugs and guide clinical decisions. The underlying question is not whether the system works (it does so with an unprecedented level of accuracy), but what ethical and regulatory protocols apply to a scenario where scientifically operational knowledge becomes computationally inaccessible to human verification.

The second is the **asymmetry of liability in emerging damages.** Legacy ethical and legal frameworks assume identifiable actors with discernible intentions. AI systems, however, can produce harms without direct deliberate intent and with causation distributed among multiple actors (designers, trainers, deployers and users) creating what the literature[174] calls the 'many hands problem': situations in which moral and legal responsibility is structurally diluted as the causal chain fragments. Current liability frameworks and compliance regimes have no structured response to this type of diffuse causation.

The third is the **representation of people unborn yet.** AI systems are trained on historical data. Their biases reflect the human past, amplified at scale. No AI ethics framework today incorporates mechanisms[175] that represent the interests of generations that have not yet been born, and thus have not produced data, yet will live with the consequences of systems designed in the present. This issue is particularly relevant for applications with long temporal horizons, from urban planning to climate risk modeling.

The fourth is **free will as a systemic risk**. An increasing proportion of individual and organizational decisions are de facto mediated by AI systems whose offerings are highly concentrated: three vendors account for 88% of global enterprise spending on language models[176], and the underlying cloud infrastructure market is similarly concentrated. This concentration implies that biases introduced deliberately or accidentally in one of these models do not affect individual decisions, but millions of simultaneous processes in different organizations, industries and countries. The cognitive diversity of a society (i.e., its ability to reach different conclusions by different paths) depends, in part, on the systems that mediate its thinking being neither homogeneous nor oligopolistic.

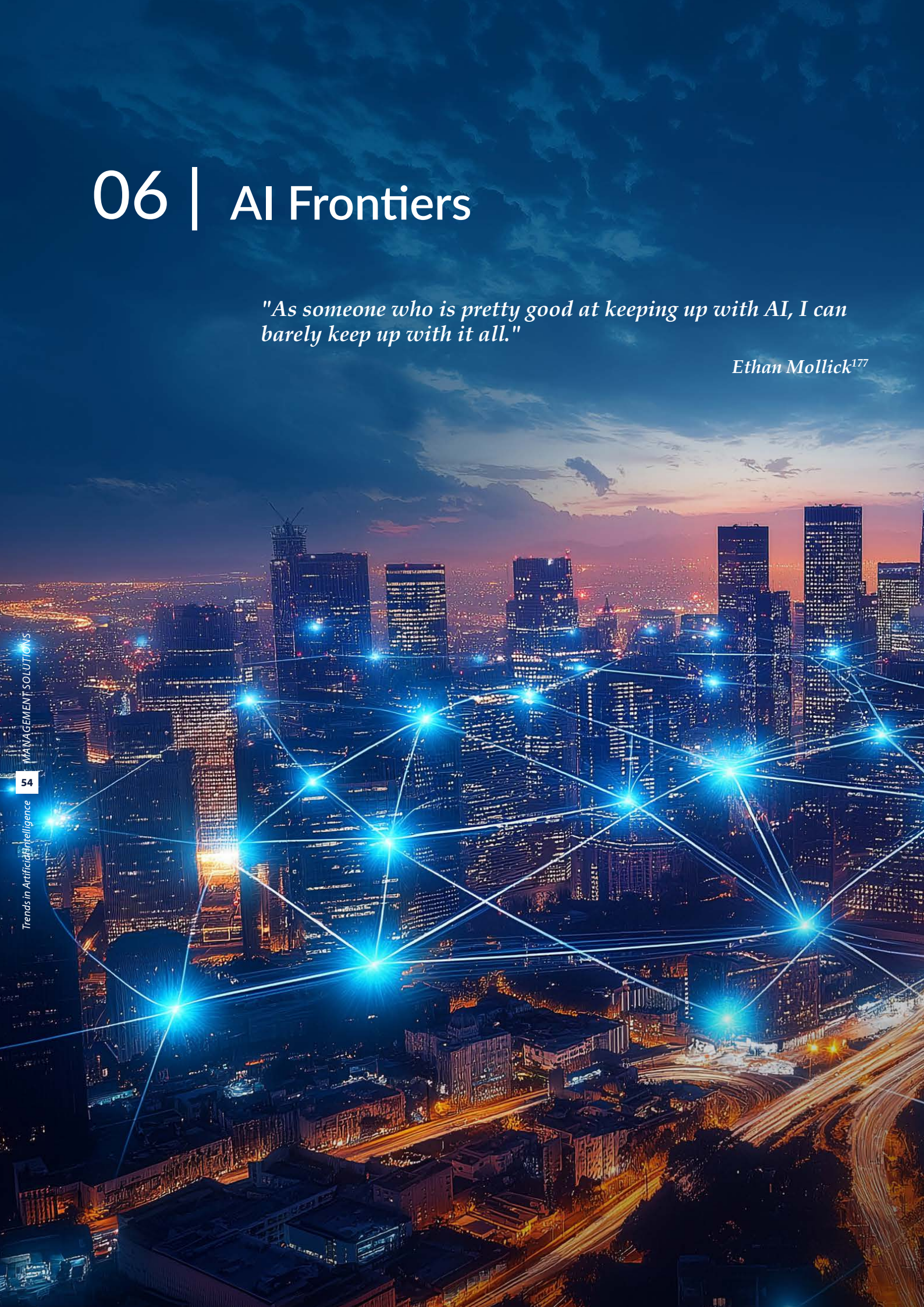
These four fractures are not arguments against the adoption of AI, but the conceptual territory that organizations that seriously manage this adoption must learn to inhabit. Their relevance is not diminished by the fact that they are not included in current regulatory frameworks; on the contrary, it is precisely because they are unaddressed that these risks are the least visible and carry the greatest potential for harm.

172 Anthropic (2026).
173 Google DeepMind (2025).
174 Thompson (1980), Nissembaum (1996).
175 Rawls (1971), Parfit (1984).
176 CEP (2026).

06 | AI Frontiers

"As someone who is pretty good at keeping up with AI, I can barely keep up with it all."

Ethan Mollick[177]



The trends discussed so far describe transformations already underway: systems in production, regulations in place, organizations adapting. This section, however, deals with a different dimension. The six trends it encompasses are not limited to the operational present; they delineate the strategic space that already shapes investment decisions, positioning and sovereignty, even if their full effects are still unfolding.

AI geopolitics is redefining alliances and dependencies. AI-first organizations are anticipating unprecedented competitive models. Digital twins and advanced simulation are transforming how we design, experiment and decide. Ambient AI is blurring the boundary between environment and computation. Convergence with quantum computing is expanding the horizon of solvable problems. And AGI has ceased to be mere academic speculation to become an explicit strategic hypothesis in the world's leading laboratories.

177 Ethan Mollick (b. 1975), associate professor at the Wharton School of the University of Pennsylvania and AI researcher, author of the New York Times bestseller Co-Intelligence. His work focuses on the impact of AI on work and education, and he was recognized by TIME Magazine as one of the most influential people in AI in 2024.

Geopolitics and Technological Sovereignty of AI

AI as a strategic state asset

AI is no longer a sectoral technology but a strategic infrastructure comparable to energy, telecommunications or financial systems. What is at stake is not only economic competitiveness: it is the ability of states to maintain autonomy in critical decisions, from defense to financial supervision, to the management of essential infrastructures. In this context, control of foundational models, advanced semiconductors, data centers and specialized talent has become a major national power factor, and industrial policies, export controls and investment strategies already reflect this new reality.

The strategic value chain: where sovereignty is at stake

Technological sovereignty in AI is not an all-or-nothing situation: there are layers to it, and dependence can arise in any of them:

- **Hardware:** TSMC produces over 90% of the world's most advanced chips, ASML is the sole global supplier of the EUV lithography needed to make them, and NVIDIA controls more than 85% of the GPU market for frontier AI model training. Three companies, three countries, one structural bottleneck.
- **Infrastructure and models:** three hyperscalers (AWS, Azure, Google Cloud) concentrate roughly two-thirds of the world's total compute capacity; on top of it, OpenAI, Anthropic and Google DeepMind develop the most capable foundational models, with cumulative advantages in data, talent and investment that are hard to replicate.
- **Talent:** frontier research remains geographically concentrated, with international mobility turning migration policy into technology policy.

No one country or organization controls all these layers simultaneously. The strategic question is not how many layers are controlled, but which are mission critical and which are manageable through diversification, agreements or redundancy.

Three competing models

The United States, China and Europe have articulated structurally distinct responses, which are not mere differences of emphasis but geopolitical projects with profound consequences for global alliances, markets and standards.

The United States combines the primacy of the private sector in model development with increasing state intervention in the supply chain: the "CHIPS and Science Act" mobilizes tens of billions in domestic semiconductors, and export controls on advanced chips to China represent the greatest technological restraint among major powers since the Cold War. The strategy is clear: maintain the edge in frontier models and deprive competitors of the hardware needed to achieve it.

China combines massive state investment, civil-military integration and an explicit strategy of technological self-sufficiency and comprehensive control of the digital ecosystem. Its stated goal is self-sufficiency across the entire value chain by 2030, from semiconductors to proprietary foundational models. U.S. export restrictions have accelerated this agenda: DeepSeek demonstrated in 2025 that Chinese innovation can produce competitive models with previous-generation hardware, complicating the logic of containment through chip control.

Europe has staked regulatory leadership as a vector of geopolitical influence. The AI Act and GDPR have generated a real "Brussels effect": global companies adapt their products to European standards because the European market is too big to ignore. However, Europe maintains significant infrastructural dependencies: its most advanced foundational models are American, its cloud is largely foreign, and its sovereign computing capacity is limited. ASML is the notable exception: the Dutch monopoly in EUV lithography makes Europe an indispensable player in the global semiconductor chain.

The rest of the world navigates between these three poles with very uneven capabilities. Some emerging countries are articulating strategies of their own with growing ambition: India is developing foundational models in local languages and negotiating its position in semiconductor supply chains; Brazil is leading AI governance initiatives in Latin America; the African Union is advancing continental frameworks for digital sovereignty. However, for most countries the choice between incompatible ecosystems remains more implicit than deliberate, with real risks of structural dependency that the literature calls "data colonialism": the mining of local data to train models that are deployed globally, without source countries capturing value or maintaining effective control over their digital infrastructure.

Fragmentation and risk of decoupling

The world is moving toward partially incompatible technology ecosystems. Chip export controls, geo-fenced data, divergent technical standards and parallel infrastructures are shaping what some analysts call "technoblocks": spheres of technological influence with their own governance, security and value logics. The Chip 4 alliance (US, Japan, Taiwan, and South Korea) coordinates Western semiconductor strategy; China cultivates its own sphere through the Digital Silk Road. A complete decoupling between the US and Chinese ecosystems would force third countries and organizations to choose, with potentially prohibitive transition costs.

The limits of absolute sovereignty

Total self-sufficiency in AI is economically unfeasible for most countries and organizations. Recreating from scratch the entire value chain (from mining critical minerals to developing foundational models) requires investments and scales that only two or three global players can sustain. Realistic technological sovereignty is not isolation: it is the effective ability to decide, diversify suppliers, negotiate terms and avoid strategic lock-in at the truly critical layers. For most countries, AI sovereignty in practice boils down to controlling the only asset over which they have effective jurisdiction: the data generated in their territory.

Implications for organizations

The geopolitical debate lands in concrete corporate decisions. Reliance on a single foundational model provider exposes organizations to risks of lock-in, changes in pricing or terms of service, and even regulatory restrictions arising from tensions between jurisdictions. Multi-jurisdictional data localization and regulatory compliance add further layers of operational complexity.

Multi-model and multi-cloud strategies, previously justified on technical performance and cost grounds, now take on an additional strategic dimension: they are the organizational equivalent of diversifying sovereign dependencies.

AI-First and AI-Only Organizations

Definition and taxonomy

The expansion of agentic systems raises a question that until a few years ago was theoretical: can an organization function with AI as its central cognitive architecture, with human labor being the exception? To answer it precisely, it is useful to distinguish three stages that business practice often confuses:

- An **AI-enhanced** organization uses AI to improve existing processes; this is the predominant model today.
- An **AI-first** organization designs its processes and structure based on AI capabilities, assigning to human judgment only those tasks where its comparative advantage is clear.
- An **AI-only** organization operates core functions entirely without human labor: No consolidated examples exist as of this writing, and its viability in regulated environments remains a working hypothesis.

The state of the art: AI-first as an operational frontier

The most advanced examples of AI-first organizations come, significantly, from the pure technology sector, where the absence of regulatory constraints on automation and the digital nature of the product allow the model to be pushed to its current limits.

Midjourney, the AI image generation platform, earned over $500 million in revenue in 2025 with a staff of approximately 163, no marketing investment and no external funding[178]. Development platform Cursor (Anysphere) reached $500 million in annual recurring revenue in May 2025, making it the fastest growing SaaS company in history with less than 50 employees[179]. The revenue per employee ratio of these companies (over $3 million in both cases) is roughly 10 times higher than historical benchmarks for the technology sector and large global banking groups.[180]

In the field of financial services, an advanced case is MYbank, a Chinese digital bank owned by Ant Group, which since 2015 has been operating under the principle of zero human intervention in SME credit approval. Its "310" model – three-minute application, one-second approval, zero human intervention – has served more than 50 million SMEs. The system uses cash flow forecasting models with over 95% accuracy and relies on satellite geolocation data for agricultural risk assessment[181]. MYbank operates without a branch network or sales force, although it maintains engineering and management teams: it is an AI-first model in its core operations, not AI-only as a whole.

Why is there still no AI-only organization?

The gap between AI-first and AI-only is not just technological; it is also regulatory, legal and organizational. In regulated sectors, such as banking and insurance, current regulations require oversight and human responsibility for material decisions. The European AI Act classifies AI applications in credit, healthcare, life

178 Midjourney (2025).
179 Sacra (2025).
180 Dealroom (2025).
181 CKGSB (2025).

insurance, and critical infrastructure as high-risk systems, with explicit requirements for human oversight[182]. Eliminating this oversight at the operational core of a financial institution would be incompatible with the European prudential framework.

In unregulated sectors, the current limit on AI-only operations is not regulatory but technical and organizational. Autonomous AI agents can execute complex tasks, but their error rate in extended workflows, inability to handle situations not contemplated in training and the absence of legal liability mechanisms equivalent to those of a legal entity mean that fully eliminating human labor from core operations would introduce operational risks that are not yet manageable.[183]

The case of Klarna illustrates the current limits: the company reduced its workforce from 7,400 to approximately 3,000 between 2022 and 2025 through hiring freezes and extensive automation, with one AI assistant managing the equivalent of 853 employees in customer service[184]. Their trajectory helps define the practical threshold between tasks that AI can perform autonomously with sufficient quality and those where human judgment continues to provide differential value today.

Predictions from industry leaders

The heads of leading frontier AI labs increasingly forecast that AI-only organizations may be on the immediate horizon. Sam Altman, CEO of OpenAI, stated in 2024: "We're going to see ten-person companies with billion-dollar valuations very soon [...] There's a bet in my chat group of executive friends about when the first one-person company valued at a billion dollars is going to exist, which would have been unimaginable without AI. And now it's going to happen"[185]. Asked in May 2025 about when that scenario would materialize, Dario Amodei, CEO of Anthropic, replied, "2026".

Amodei develops the argument in his January 2026 paper, where he describes the functional equivalent of "a country of geniuses in a data center," i.e., 50 million agents more capable than any Nobel Laureate, operating at between ten and one hundred times human speed; and he estimates that 50% of entry-level jobs could be disrupted within one to five years[186]. The same paper notes that Anthropic already runs most of the code it produces using AI, approaching full operational autonomy in software development.

Can existing organizations become AI-only?

The underlying strategic question is not whether AI-only organizations will exist, but how they will come to exist. The answer is counterintuitive: they are unlikely to emerge from the transformation of existing organizations. Clayton Christensen documented[187] in The Innovator's Dilemma that incumbent companies are structurally incapable of adopting disruptive technologies from within: their processes, incentives and customer bases are optimized for the incumbent model, and any internal disruptive initiatives compete at a permanent disadvantage for resources and management attention. The transition to AI-first exacerbates this logic: an organization with tens of thousands of employees has its processes designed for that human scale. Those processes are not redesigned; they are replaced.

The pattern emerging in Asia points to an alternative path: the creation of new entities, with their own brand and no operational heritage, that compete freely until they reach critical mass and cannibalize the original company. Ping An, the world's largest insurer by premiums written, incubated between 2013 and 2022 eleven independent technology subsidiaries (including OneConnect, Lufax and Ping An Good Doctor), five of which were listed as standalone entities[188]. DBS Bank created Digibank as a separate digital bank operating with one-fifth of the resources per customer of a conventional bank, with its lessons feeding back into the parent organization's architecture[189]. The mechanism is identical: a new, non-legacy entity that scales without the constraints of the parent organization and, if the experiment fails, is closed without dragging down the original company.

In Europe and, to a lesser extent, in the United States, this mechanism encounters structural frictions that go beyond AI regulation. In Europe, the introduction of workplace AI systems may require consultation or negotiation with works councils, and in some countries their explicit agreement[190]. Collective bargaining agreements in employment-intensive sectors incorporate clauses limiting automation. Employee data protection under GDPR adds additional complexity[191]. The result is an asymmetry with unintended strategic consequences: western regulation makes it difficult for existing organizations to build the AI-first entities that would eventually challenge them.

Once technology reaches the point where an AI-only organization is viable, the question of who will build it first will likely come down to geography.

182 AI Act (2024).
183 Amodei (2026).
184 Fortune (2025c).
185 Altman (2024a).
186 Amodei (2026).
187 Christensen (1997).
188 IMD (2023).
189 DBS (2024).
190 Baker McKenzie (2025).
191 ILO (2025c).

Digital Twins and Simulation of Human Behavior

From Aerospace Engineering to Universal Simulation

The concept of the digital twin has a precise date and place of birth. In October 2002, Michael Grieves presented at a forum of the Society of Manufacturing Engineers what he called "Conceptual Ideal for Product Lifecycle Management": the idea that any physical object could have a digital correlate that would dynamically represent it throughout its life cycle, synchronizing in real time the state of the real object with its virtual representation. The term "digital twin" was later coined by John Vickers, NASA's chief engineer, who formalized the concept in the agency's 2010 technology roadmap[192]. NASA's definition in that document remains the most accurate available: "a multiphysics, multiscale, probabilistic simulation of a vehicle or system that uses the best available physical models, sensor updates, and fleet history to mirror the life of its physical twin."

The starting point is important because it reveals the implicit premise that has guided the development of digital twins for two decades: a digital twin works well when the system it models obeys known, deterministic physical laws. A gas turbine, an aircraft fuselage, an electrical grid: complicated systems with many components, but in principle fully modelable if sufficient computational power and sensor data are available. Under that premise, the technology matured steadily. Today, digital twins of physical assets are operational in fields such as advanced manufacturing, energy, infrastructure and aviation, delivering measurable reductions in unplanned maintenance times and significantly accelerating product design cycles.

The epistemological boundary: complicated systems versus complex systems

Expanding the concept beyond the physical domain has revealed a limit that is not technological but epistemological. Michael Batty, the most recognized academic authority on computational modeling of cities, explains it precisely: digital twins work in complicated systems (many parts, but determinable behavior in principle) and encounter structural difficulties in complex systems, where the overall behavior emerges from the interaction of agents and cannot be deduced from the properties of their individual components193. A city, an economy, a financial market, or a human organization are complex systems in this precise technical sense.

Philosopher Stefano Moroni expresses the argument in even more direct terms: the limitations of urban digital twins are not temporary (they will not disappear with more data or greater computational power), but instead arise from the intrinsically emergent nature of social systems[194]. The unpredictability of detail in a complex system is not an information deficit; it is a property of the system. This has immediate practical implications: a digital twin of a manufacturing plant can reliably predict when a bearing will fail; a digital twin of a city can approximate aggregate traffic trends, but cannot reliably predict the effect of a housing policy on ten-year residential segregation patterns. The distinction is not one of degree, but of nature.

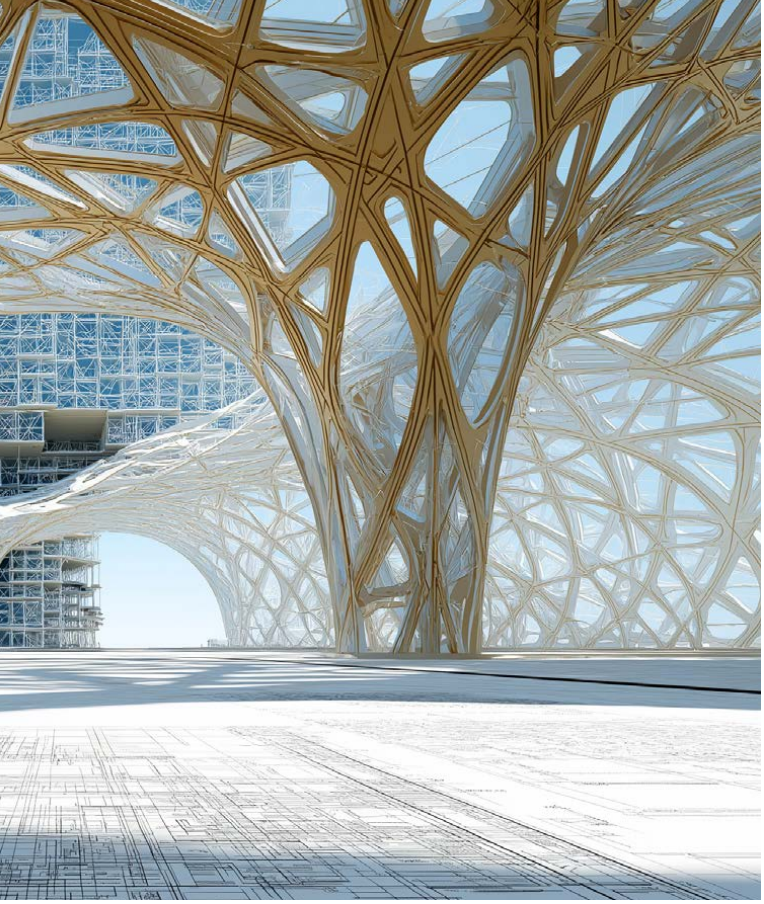
This epistemological boundary has defined the ceiling of the digital twin field for decades. It is precisely here that large-scale language models introduce a discontinuity that deserves attention.

The tipping point: human behavior becomes modelable

In April 2023, a team of Stanford researchers published a paper that launched a radically new line of work. Joon Sung Park and his coauthors created 25 computational agents (each endowed with an identity, a persistent memory, a set of social relationships, and a reasoning ability based on a language model), and placed them in a simulated environment equivalent to a small city[195]. The agents woke up, ate breakfast, went to work, formed opinions, initiated conversations and coordinated collective activities without these behaviors having been explicitly programmed: they emerged from the interaction between each agent's individual memory, their ability to reflect on past experiences and their model of the social environment. The paper won the Best Paper Award at the 2023 ACM Symposium on User Interface Software and Technology. The scientific community recognized that something qualitatively new had happened.

192 Grieves-Vickers (2017).
193 Batty (2024).
194 Moroni (2025).

195 Park (2023).



The underlying reason is that large-scale language models have absorbed, during training, an extraordinary amount of recorded human behavior: conversations, decisions, reasoning patterns, emotional responses, implicit social norms. They have not learned the laws of human behavior explicitly (no one knows them with that precision), but they have developed a statistically dense approximation that, under controlled conditions, generates plausible behaviors. For the first time, the premise that prevented modeling complex social systems has been partially lifted: not because human behavior has ceased to be emergent, but because there is now a behavior generator rich enough to populate a simulation with credible agents.

The natural extension of this work came in November 2024. The same Stanford team published the results of a different scale experiment: 1,052 real people, interviewed in depth about their lives, attitudes, and experiences, were turned into agents that replicate their responses and behaviors in standardized surveys and social experiments. The generative agents replicated real individuals' responses on the General Social Survey with 85% accuracy (statistically comparable to the individual's own natural variability when answering the same survey two weeks later) and produced comparable results in replicating personality traits and in social science experiments[196]. What began as a conceptual demonstration with fictitious characters in 2023 became, by 2024, an empirically validated methodology with real people.

196 Park (2024).

Current applications and future horizon

The implications of this leap are cross-sectoral. In market research, the startup Simile – founded by Joon Sung Park together with Michael Bernstein and Percy Liang, the co-authors of the founding paper, and backed in February 2026 with $100 million by Index Ventures with participation from Fei-Fei Li and Andrej Karpathy, builds digital twins of real people to help companies simulate their customers' behavior before launching a product, modifying a pricing policy or redesigning a user experience[197]. In a public demonstration, the platform correctly predicted eight out of ten answers to the questions asked by analysts in a simulated call[198]. The global market research industry, valued at $142 billion[199], is facing structural disruption: what today requires weeks of fieldwork can be executed in hours on synthetic populations.

In public policy and urban planning, a more nuanced but equally transformative use is emerging: not the digital twin as a predictive oracle, but as a scenario laboratory where the consequences of different interventions can be explored before committing real resources[200]. In financial regulation, this approach has direct application in the stress testing of adverse macroeconomic scenarios and in the simulation of market behavior in the face of regulatory interventions.

The trajectory of this domain suggests a qualitatively different scale ahead. If today it is possible to simulate a thousand real people with high fidelity, the question on the immediate horizon is what happens when that number reaches a million, a hundred million, an entire society modeled in real time. The applications in public policy, economic regulation and institutional design will be of a different order of magnitude than market research: not anticipating what product a consumer will buy, but predicting how a population will respond to a tax reform, a health crisis or a change in monetary policy before that intervention is implemented in the real world. This capability has no historical precedent, nor, as yet, governance framework to regulate it.

197 Index Ventures (2026).
198 Bloomberg (2026).
199 ESOMAR (2024).
200 Bettencourt (2024).

Ambient AI and Invisible Computing

The interface is the environment

Ambient AI – or ambient intelligence – is AI that operates without being invoked. Unlike conventional systems, which respond to an explicit instruction from the user, ambient systems continuously observe the context, infer needs and act proactively. The interface disappears not because it has been improved, but because the system no longer needs it: the environment itself becomes the point of interaction. Computing becomes "invisible" in the literal sense: embedded in objects, spaces and processes without the user perceiving it as such.[201]

This inversion (from the user going to the system to the system coming to the user) is made possible today by the convergence of three simultaneous developments: the miniaturization of models capable of running on edge devices while maintaining a continuous state (locally updated cumulative user memory) without relying on cloud connectivity (edge AI and TinyML), the densification of physical and biometric sensor networks, and the ability of LLMs to reason about heterogeneous and ambiguous context in real time[202]. None of the three is new on its own; their simultaneous maturity is what makes Ambient AI capable of moving from concept to operational deployment.

Current state of deployment

The most documented example of Ambient AI in operation are ambient AI scribes in clinical settings: systems that continuously listen to the patient-physician conversation, infer the clinical context without explicit instruction, and automatically generate encounter documentation. A UCLA randomized clinical trial evaluated two platforms – Microsoft DAX and Nabla – across 238 physicians from 14 specialties and over 72,000 encounters: it reduced documentation burden and improved indicators of professional burnout.[203]

The system was not invoked iteratively during the consultation: it listened, inferred, wrote. It is still a bounded form of ambient intelligence (limited context, clear purpose, defined episode). Mature Ambient AI will operate beyond individual consultations, at the scale of the entire hospital, correlating longitudinal patterns with no user-determined start or end point.

The physical and digital future

Today's deployments are the tip of a broader transformation. In the coming years, Ambient AI will extend to physical and digital environments and make current cases look rudimentary:

Physical environments

- **Adaptive workspaces.** The environment infers the occupant's attentional state from heart rate, rhythm variability, movement patterns, and reconfigures temperature, light, and noise level to optimize cognitive performance without conscious user intervention.

201 Bimpas (2024); Hernández-Torres (2025).
202 Heydari (2025).
203 Lukac (2025).

- **Anticipatory industrial maintenance.** Systems will not alert when equipment fails: they will detect the behavioral pattern preceding the failure early enough to reorganize production. The disruptive event disappears from the operating horizon.
- **Individually profiled wearables.** Next-generation devices will not compare the wearer's vital signs to population averages, but to his or her own physiological history. The alert will be triggered before the symptom becomes conscious to the wearer.
- **Reactive urban infrastructure.** Transportation, lighting, and waste management networks that self-adjust in real time to inferred usage patterns, without explicit centralized planning or human intervention in the loop.
- **Invisible home care.** Systems that continuously monitor elderly people or those with chronic conditions, detect anomalies in routines (sleep patterns, mobility, feeding) and activate alert or intervention protocols without the user having requested anything.

Digital environments

- **Development environments that anticipate the problem.** Proactive programming assistants will evolve into systems that, before the developer identifies the error, will have mapped out the likely solution space and presented options at the cognitively appropriate time.[204]
- **Attention management, not just information management.** Systems will not serve information when it is available, but when the user is in a position to process it: modeling attentional state throughout the day and calibrating the timing of interruption.
- **Continuous organizational context.** Systems will know at all times the status of projects, pending communications and ongoing decisions, and will surface relevant information to each team member without being requested.
- **Autonomous resource negotiation.** Environmental agents will manage tasks on behalf of the user (scheduling, budgeting, arranging access to services) within defined parameters, without requiring explicit approval for every low complexity decision.

Implications that technology does not solve

Ambient AI does not only raise privacy questions; it introduces a broader set of tensions that current governance frameworks have not resolved.

The first is the nature of error. In an invoked system, error is visible: the user asked for something, the system responded badly. In an ambient system, the error may not be perceived because there was no explicit request against which to compare the response. The UCLA **ambient scribe** recorded clinically significant inaccuracies in a proportion of encounters[205]: in an invisible system, the error detection mechanism has to be deliberately designed, because it does not emerge naturally from the interaction.

The second is the asymmetry of power between those who design the environment and those who inhabit it. In a hospital, an office or a public building, users do not choose whether the environment is intelligent: they inhabit a space whose inferences about their behavior have been configured by a third party. Tshilidzi Marwala, Chancellor of the United Nations University, puts it precisely: Ambient AI has an appetite for data – intimate, behavioral, biometric – that renders conventional notions of informed consent structurally inadequate[206]. The European AI Act, designed for invoked systems with bounded functions, does not provide a satisfactory response to these continuous observation environments.

The third is cognitive dependency. A system that proactively manages the user's attention, information flow and interruptions not only assists its work: it shapes its cognitive architecture. The question posed in 2003, "is context-aware computing taking control away from the user?"[207] has gone unanswered for decades. The scale at which Ambient AI poses it today turns what was an academic question into a design issue with immediate operational consequences.

The fourth tension is causal accountability. In invoked systems, traceability is relatively straightforward: there is an instruction, a response, an attributable decision moment. In environmental systems, the causal chain is blurred. If an anticipatory maintenance system reorganizes production and that reorganization conditions subsequent human decisions, the boundary between technical agency and human agency is not clear. Current regulation, including the AI Act, assumes predictable finality and ex ante risk assessment; Ambient AI introduces emergent finality and continuous adaptive behavior, directly challenging existing compliance mechanisms.

Ambient AI doesn't just change how we work or how we take care of ourselves: it changes the sequence between need and awareness. A system can know what we need before we do. The possibility that this capability may be deployed at the scale that the field's trajectory suggests raises questions that reach beyond technology and regulation to something more fundamental: what it means to make our own decisions in an environment that has already anticipated them.

204 Chen (2025); Pu (2025).

205 Lukac (2025).
206 Marwala (2025).
207 Barkhuus (2003).

Interaction between AI and Quantum Computing

Two distinct technologies, an intersection that matters

AI and quantum computing are independent technologies with completely different principles, time horizons and use cases. AI is already operational on an industrial scale; quantum computing is still, for the most part, an advanced research field with very limited deployments. Their interaction, in both directions, has concrete implications for any organization that relies on digital systems.

A classical computer solves problems by testing options one at a time or in parallel, but always within a space of possibilities that grows in a manageable way. Some problems exceed this capacity: optimizations with thousands of interdependent variables, simulations of molecular systems, or certain mathematical problems that form the basis of modern cryptography. A quantum computer works in a fundamentally different way: instead of trying options one by one, it can simultaneously explore a space of possibilities of a dimensionality that no classical system can represent. For this specific class of problems – not all problems – the performance difference is not incremental, but many times greater than what classical computers can achieve.

The problem is that building a quantum computer that works reliably has proven extraordinarily difficult. Quantum information is extremely sensitive to environmental perturbations – temperature, vibrations, electromagnetic interference – and errors accumulate rapidly. For decades, the field advanced much faster in theory than in hardware. That partially changed in December 2024.

The 2024 milestone and what it means

Google published results from its Willow processor in Nature. Willow is the first system to show that as more computational components are added, errors decrease rather than increase[208]. This outcome had been theoretically predicted since 1995, but no system had previously managed to realize it. The significance lies not in the performance figures (which are impressive though based on artificial benchmarks), but in what it means for the trajectory of the field: the obstacle that for thirty years had prevented scaling these systems reliably has been overcome in the laboratory.

The gap between that achievement and a quantum computer with commercial applications remains considerable. Google's own researchers put that horizon at around the end of the decade. But the direction is no longer in dispute: the central problem was in error correction, and that problem now has a proven solution. What remains is engineering for scale, not a scientific leap in a vacuum.

Three ways this affects AI

The intersection of quantum computing and AI operates on three distinct planes, with different urgencies.

The first is the **acceleration of machine learning**. Training a large-scale AI model is essentially a mathematical optimization problem: finding the values of billions of parameters that minimize prediction error across an enormous search space. This is precisely the kind of problem where quantum computing offers a theoretical advantage. It has been formally demonstrated[209] that fault-tolerant quantum systems could substantially speed up the training of large AI models, reducing both computational time and energy use. Achieving this requires hardware that does not yet exist at sufficient scale. When it does becomes available, however, it could radically alter the economics of AI model training, which is currently dominated by organizations able to afford massive-scale GPU infrastructure.

The second plane is **quantum machine learning** itself: using quantum processors to run ML algorithms more efficiently. Here the literature is more cautious and describes both the promises and the actual obstacles. Quantum computers do not automatically outperform classical systems in all learning tasks, and the advantage they offer is far from universal. In many cases, classical systems with access to data can perform as well as quantum systems, even on problems specifically designed to favor quantum approaches. In other words: data, used well, can offset the quantum advantage in many situations[210]. The hype about quantum AI as a universal accelerator is ahead of the evidence; real applications will be problem-specific rather than broadly transformative.

The third plane flips the perspective on impact: **instead of quantum computing serving AI, it poses a threat** to the security infrastructure that underpins all digital systems, including AI systems. All the cryptography that protects digital communications today (banking transactions, identity authentication, secure channels between systems) rests on mathematical problems whose difficulty is assumed to be unassailable for classical computers. A sufficiently powerful quantum computer would solve them directly. This plane is the most urgent, because some of its effects are already emerging.

The threat that is already active

The strategy known as **"**harvest now, decrypt later**"** consists of capturing encrypted communications today with the intention of decrypting them when quantum computing matures sufficiently. State actors with advanced intelligence capabilities have been applying it for years[211]. Data requiring confidentiality for decades (medical records, trade secrets, regulatory communications or sensitive financial information) are being compromised now, regardless of when the quantum system capable of decrypting them arrives.

208 Google Quantum AI (2024).

209 Liu (2024).
210 Huang (2021).
211 Mascelli (2025).

The most rigorous institutional response is that of the US NIST, which in August 2024 – after eight years of work and more than eighty proposals from research teams around the world – published the first three standards for quantum-resistant cryptography[212]. The new algorithms are based on mathematical structures for which no efficient quantum attack is known. NIST urges organizations to begin migration immediately; the deadline for U.S. federal systems is 2035. For financial entities with long-lived data, that deadline is not the point for starting the transition: it is the final limit for completing the migration.

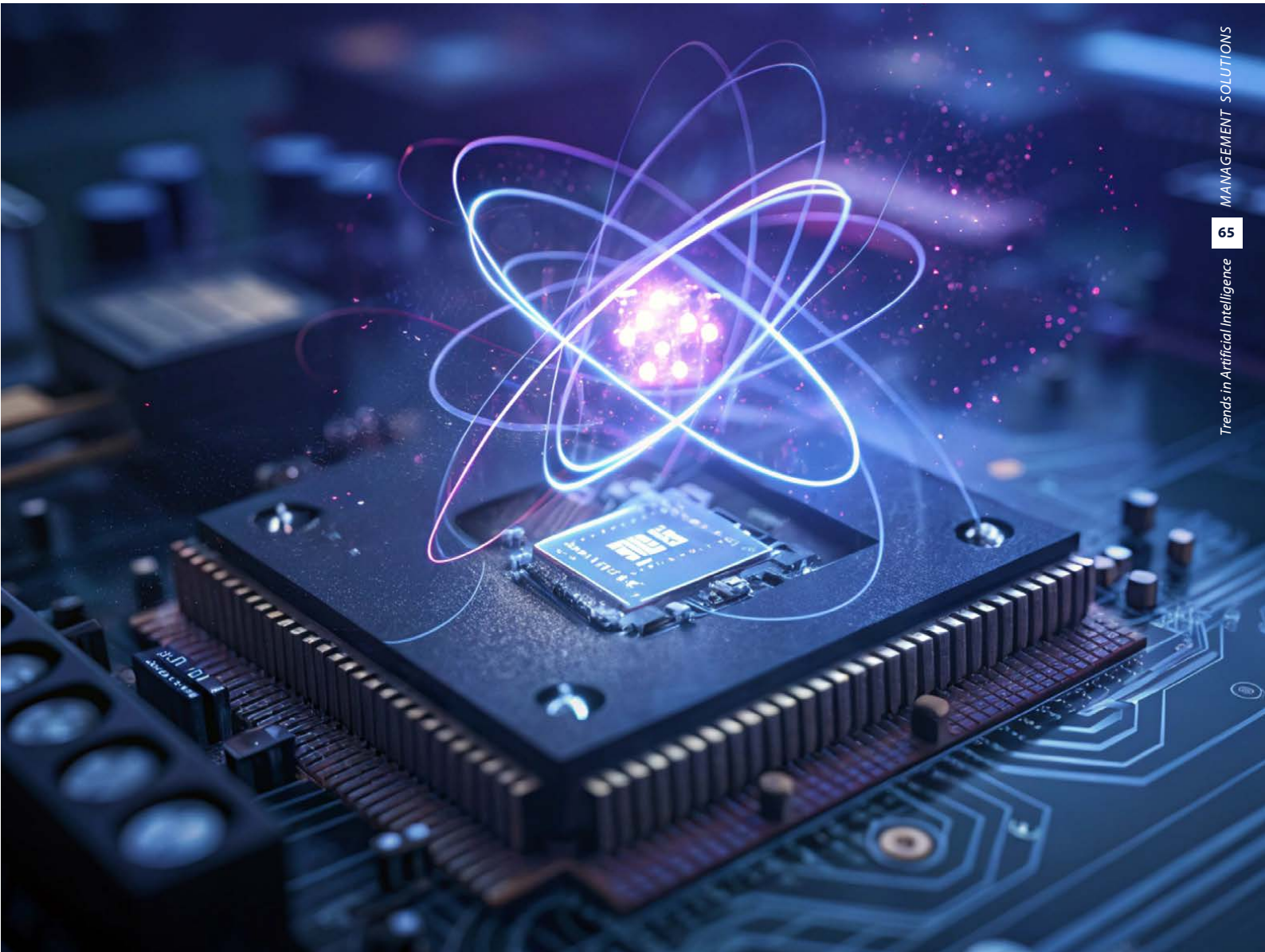
The horizon of the next decade

Over the next few years, the intersection between AI and quantum computing will shift from being a technology foresight topic to an element with operational impact on two simultaneous fronts.

The **offensive front** (the acceleration of AI capabilities) will come with the maturity of fault-tolerant hardware, predictably towards the end of the decade. Initial access will be via cloud services, replicating the trajectory that high-performance computing followed with GPUs: first accessible only to the best-funded, then democratized through competition among vendors. Organizations that have developed AI competencies by then will be better positioned to take advantage of that acceleration when it arrives.

The **defensive front** (cryptographic migration) admits no delay. The readiness window narrows as quantum processors scale, and cryptographic infrastructure migration in complex organizations takes years. Inventorying vulnerable assets, prioritizing by data lifetime, and transition planning are tasks that must begin now, not when the relevant quantum computer is operational and it is too late.

212 NIST (2024b).



Artificial General Intelligence (AGI) as a Strategic Horizon

What is AGI, and has it arrived yet?

The term "artificial general intelligence" (AGI) first appeared in 1997, at a nanotechnology conference in Palo Alto. Its author, Mark Gubrud, a doctoral candidate at the University of Maryland, was not describing a desirable technological goal: he was warning of a risk. In his paper "Nanotechnology and International Security"[213] , Gubrud defined AGI as systems capable of "rivaling or surpassing the human brain in complexity and speed, acquiring, manipulating and reasoning with general knowledge, and being usable in any phase of operations where human intelligence would be required." The definition went unnoticed for nearly a decade, until Ben Goertzel and Shane Legg rescued it and popularized it as a technical label in titling their collective book Artificial General Intelligence[214]. The term had been born as a cautionary tale, but became a mission.

In February 2026, Nature published almost simultaneously two texts crystallizing arguably the most relevant debate in contemporary technology. The first, signed by four UC San Diego academics[215] - specializing in philosophy, machine learning, linguistics and cognitive science – states unambiguously that AGI already exists: today's LLMs pass the Turing test, win gold medals in mathematical Olympiads and collaborate with humans in proving theorems. The second[216], published a fortnight later as correspondence in the same journal, replies that this conclusion is only possible by redefining the concept beyond recognition: the classical definition of AGI formulated in 2007 requires robustness under novelty, transferable generalization and goal autonomy, and current systems do not meet these requirements. The fact that top researchers, with access to the same systems and data, reach opposite conclusions reflects that "general intelligence" is a continuous concept without precise thresholds.

Fei-Fei Li herself, who built the foundation for modern computer vision and works in spatial intelligence, bluntly admits as much: **"I struggle with this definition of AGI, to be honest"[217]**. Dario Amodei, CEO of Anthropic, goes further: he openly declares that he does not like the term, preferring to speak of "powerful AI": systems with intellectual capabilities comparable or superior to those of a Nobel Prize winner in most disciplines.[218]

That is the right perspective for organizations. The strategically relevant question is not philosophical – have we reached AGI – but functional: when can a system fully autonomously perform complete cycles of high-cognitive value work in all domains? That threshold has already been crossed in several sectors.

The current state: simultaneous brilliance and fragility

Andrej Karpathy, co-founder of OpenAI and former head of AI at Tesla, coined the concept of **"jagged intelligence"** to describe the current condition of LLMs: systems that solve mathematical Olympiad problems and fail to determine which number is higher, 9.11 or 9.9; that are fluent in dozens of languages and suffer from what he calls "anterograde amnesia"[219]: inability to consolidate learning between sessions. Fei-Fei Li describes them[220] as "wordsmiths in the dark: eloquent but inexperienced, knowledgeable but ungrounded": eloquent but without embodied experience in the physical world.

And yet these same systems already write contracts, analyze credit risk, synthesize scientific literature, generate and audit complex code, or produce regulatory summaries. They do this autonomously, at a speed and scale that no human team can match. The question is no longer when that capability will arrive; it is what we do with what has already arrived and how we prepare for what comes next.

The causal chain: what's next

The ongoing transition follows a logic of cumulative escalation. The first stage is from tool to agent: systems move beyond responding to instructions and instead pursue goals through autonomous cycles of action, observation and correction. The second stage is from agent to environmental infrastructure. As Karpathy accurately puts it[221]: LLMs are the new electricity. AI ceases to be a technology that we "adopt" and becomes a technology that "happens to us": invisible operating fabric of the systems we use, a condition of the environment rather than a tool in it.

The third stage is the one most underestimated by conventional analysis: the recursive loop. Models are already used to improve other models – generating synthetic training data, optimizing architectures, or producing research hypotheses. This creates a dynamic in which the speed of AI improvement depends on AI intelligence itself. Amodei calls it "the end of the exponential": not the point where the curve flattens, but where the acceleration becomes over-exponential, because the accelerating agent is the system being accelerated.[222]

The structural consequence of this loop is unprecedented in the history of civilization: the upper limit of reasoning available on the planet has been, since the first hominids, human intelligence. That limit is being displaced in specific domains at this time. When the shift becomes general and robust, the rate of technological and scientific change will become partially decoupled from the human capacity for understanding and verification. This is not a doomsday projection, but a structural description of what this recursive loop implies.

213 Gubrud (1997).
214 Goertzel, Legg (2007).
215 Chen, Belkin, Bergen, Danks (2026).
216 Quattrociocchi, Capraro, Marcus (2026).
217 Li (2025).
218 Amodei (2024b).
219 Karpathy (2025).
220 Li (2025).
221 Karpathy (2025).
222 Amodei (2024a).

The absorption gap

Two curves advance at radically different speeds. The technical curve – exponential, self-accelerating through the recursive loop – compresses into years what used to take decades. The organizational absorption curve – process redesign, role reconversion, governance infrastructure building, institutional change management – advances more slowly, with considerable friction: legacy systems, organizational difficulties, cultural resistance, late-arriving regulation, or shortage of talent capable of integrating these capabilities into real operations.

The gap between the two curves is the determining variable of the next decade. Competitive advantage will not come from access to the best models, which will become progressively commoditized, nor from their cost, which will continue to fall exponentially. It will come from the speed and rigor with which an organization is able to redesign itself to operate with autonomous agents effectively and responsibly. This pattern is empirically documented[223]: the most significant productivity gains do not appear where AI replaces tasks, but where it reorganizes entire processes and redefines human-machine collaboration.

Redefining the human role

The right question, therefore, is not which jobs will disappear, but what does a human do that an AI system cannot do even though it is faster, cheaper, and more consistent. The usual answer (creativity, empathy, leadership) is true but insufficient. There are dimensions that conventional analysis underestimates:

- Responsibility with real consequences. AI systems cannot be taken to court or lose a reputation built over decades. In financial, healthcare, legal and regulatory environments, human presence is a structural requirement.
- Certification and public faith. The notary who certifies, the doctor who signs, the auditor who countersigns: these acts are worthy not because of the information processing involved, but because of the personal and institutional responsibility behind them.
- Interpersonal relationship. There are contexts where what is needed is not the right answer but the presence of another person: grief, conflict, care. Replacing that presence with a more efficient system does not solve the problem.
- Formulation of worthwhile questions and moral judgment. When AI performs well at what it is asked to do, the value shifts to those who decide what to ask: who sets the goals, who identifies which problems deserve attention or who decides when there are conflicting values.
- Democratic legitimacy. Decisions affecting communities require deliberation and accountability that cannot be delegated to opaque systems, however precise they may be.

What these dimensions have in common points to something deeper than a distribution of tasks. Throughout history, humans have simultaneously been the agents who act and the subjects who are accountable for the outcomes of those actions. AI systems structurally dissociate this unity between action and responsibility. What is redefined, then, is not just what we do, but our position in the causal chain: less in execution, increasingly in intention, judgment, and accountability.

The real systemic risk: concentration

The emerging systemic risk is not that of the machine that rebels. It is that of the unprecedented concentration of cognitive capacity in a small number of actors - laboratories, corporations, states - whose advantage is self-amplifying by the same recursive loop that accelerates overall progress. Amodei writes that, if an authoritarian state were to achieve, thanks to AI, offensive dominance in cybersecurity or biology before the rest, the geopolitical consequences would be asymmetric and irreversible. Hassabis proposes, as a response, a CERN-inspired model of international collaboration: multilateral governance of the final steps towards general AI systems.[224]

The institutional response – regulatory, corporate, international – is currently far behind the speed of the problem. AGI as a strategic horizon does not require organizations to resolve the philosophical debate over whether it has arrived. It requires that they act with the awareness that its consequences are already unfolding: in the systems they operate today, in the work cycles being redefined today, and in the governance decisions that are being made – or circumvented – today.

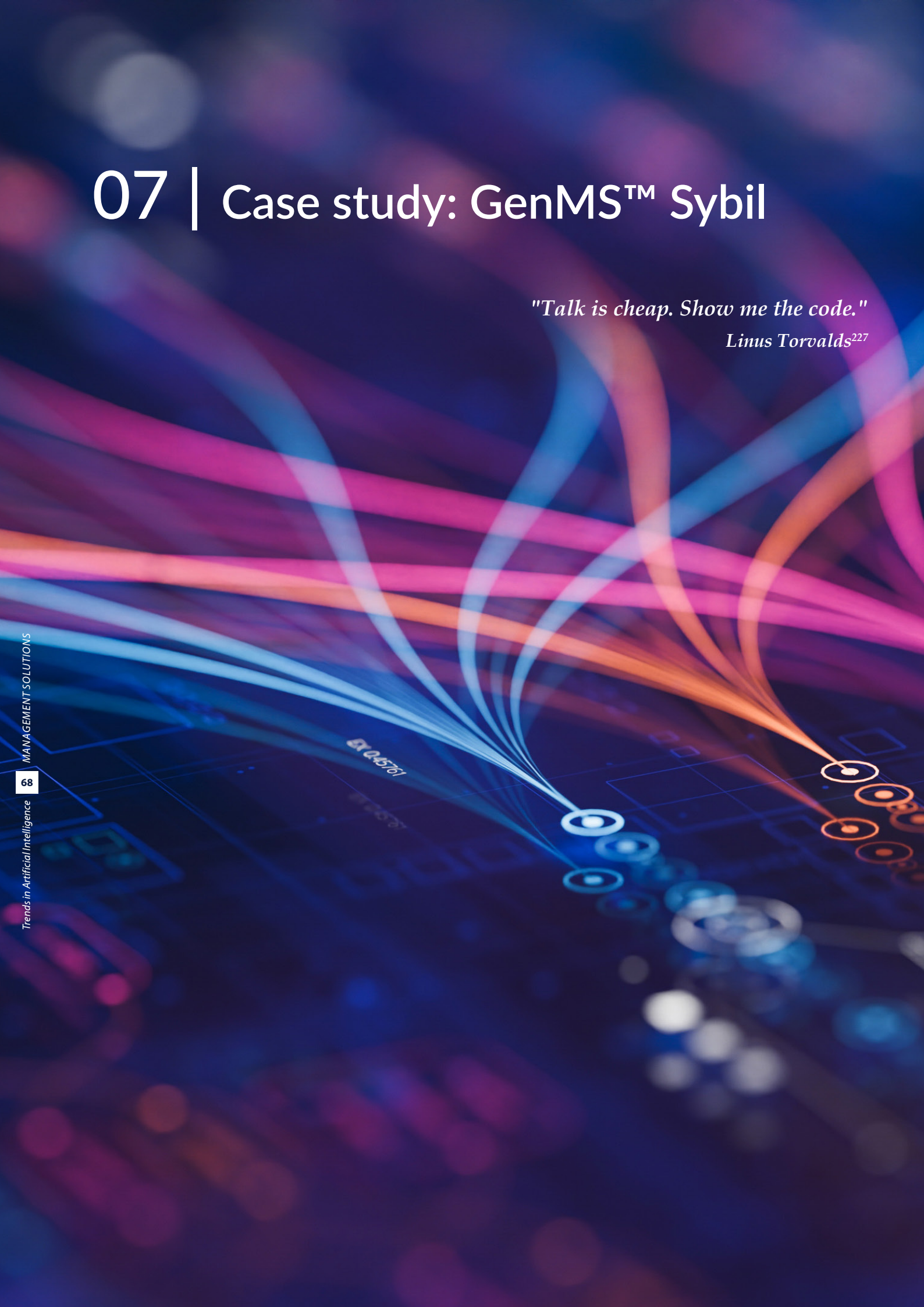
223 Mollick (2024).

224 Amodei, Hassabis (2026).

07 | Case study: GenMS™ Sybil

"Talk is cheap. Show me the code."

Linus Torvalds[227]



GenMS™ Sybil was specified, built, secured, validated and deployed in a single day's work. This case documents how[226].

225 Linus Torvalds (b. 1969), Finnish-American software engineer, creator of the Linux kernel and Git, two of the most influential free software projects in history.

226 The scope is deliberately constrained: a closed-corpus conversational assistant, with no integrations into corporate systems, no session persistence, and a limited attack surface. This constraint is not a simplification; it is a design choice. The statement does not suggest that more complex systems require equivalent effort: complexity and development time can scale non-linearly with the number of integrations, concurrent users, regulatory requirements, and operational criticality.

The system

GenMS™ Sybil is a publicly accessible conversational assistant, built on the full content of this document. It answers questions, explores implications and accompanies reflection on the trends discussed here. It does not store users' personal data and therefore does not present any difficulties with respect to the General Data Protection Regulation. The system is compliant-by-design: regulatory classification, AI Act requirements and privacy obligations were not incorporated as a subsequent layer of compliance, but as design criteria from the first specification phase.

The process: LLMOps lifecycle

The build followed the LLMOps lifecycle phases sequentially and without exception.

Data preparation. The corpus of GenMS™ Sybil is this document. The decision not to extend the system with the full content of cited sources was deliberate: doing so would have introduced copyright and intellectual property risks that are difficult to manage. GenMS™ Sybil is aware of the references, cites and links to them, but does not reproduce their content. Source control and minimization are, here, simultaneously a technical decision and a compliance requirement.

Experimentation and development. This phase produced the complete specification of the system: architecture, expected behavior, taxonomy of use cases, operational limits, quality criteria and security requirements. The specification, dozens of pages long, was built in dialogue with an LLM through vibe coding: the professional formulated objectives, evaluated proposals and made decisions; the machine materialized the intent into production technical documentation. Alternative model configurations were evaluated, prompt versions were managed from the beginning, and qualitative evaluation metrics were defined: coherence, factuality, contextual appropriateness, behavior in the face of out-of-scope questions.

Validation. The evaluation of GenMS™ Sybil integrated human review into the process, semantic stress testing and red-teaming exercises aimed at identifying undesired behavior. Validation was not a one-time event at the end of the process; it was continuous throughout the cycle. GenMS™ Atlas - Management Solutions' system for testing LLM-based systems - evaluated the system on several of its 26 dimensions: bias, consistency, privacy, robustness, explainability and regulatory compliance. Detected issues were addressed prior to deployment; those that persist are documented.

Deployment. The system build was executed by Claude Code from the full specification. The result was a consistent application, with context logic, session management and user interface. The code was audited for vulnerabilities and potential attack vectors, and the corresponding fixes were incorporated within the same development cycle. Deployment took into account the infrastructure, latency and cost implications of a generative system in production from the outset.

Monitoring. GenMS™ Sybil operates with active monitoring of costs per token, full traceability of interactions for auditing and regulatory oversight, and alerts for anomalous behavior or unanticipated usage patterns. The construction process was iterative: the first version was not the final version. Controlled iteration, with explicit evaluation criteria at each cycle, is what distinguishes industrialization from experimentation.

Architectural decisions

The design of GenMS™ Sybil involved concrete technical dilemmas, solved with explicit criteria:

- **RAG vs. full context:** full context. The underlying model has an input window of one million tokens; the document fits in its entirety in each conversation. RAG fragmentation would destroy the overall coherence that the most valuable questions require, and the cost argument that historically justified it no longer compensates for this deterioration in quality.
- **Extension of the corpus with cited sources:** discarded. The risk of copyright and intellectual property infringement is incompatible with a public access system. GenMS™ Sybil cites and links references; it does not reproduce their content.
- **Fine-tuning vs. prompting:** prompting with document in context. Fine-tuning is costly, slow and opaque to document updates. Prompting guarantees full traceability of each behavioral change.
- **Proprietary vs. open source model:** proprietary frontier model. The maturity of open source models is insufficient to guarantee the consistency and security controls required in a public system without continuous human supervision.
- **Persistent memory vs. independent sessions:** independent sessions. Data minimization, structural GDPR compliance and elimination of risk of contamination between conversations.

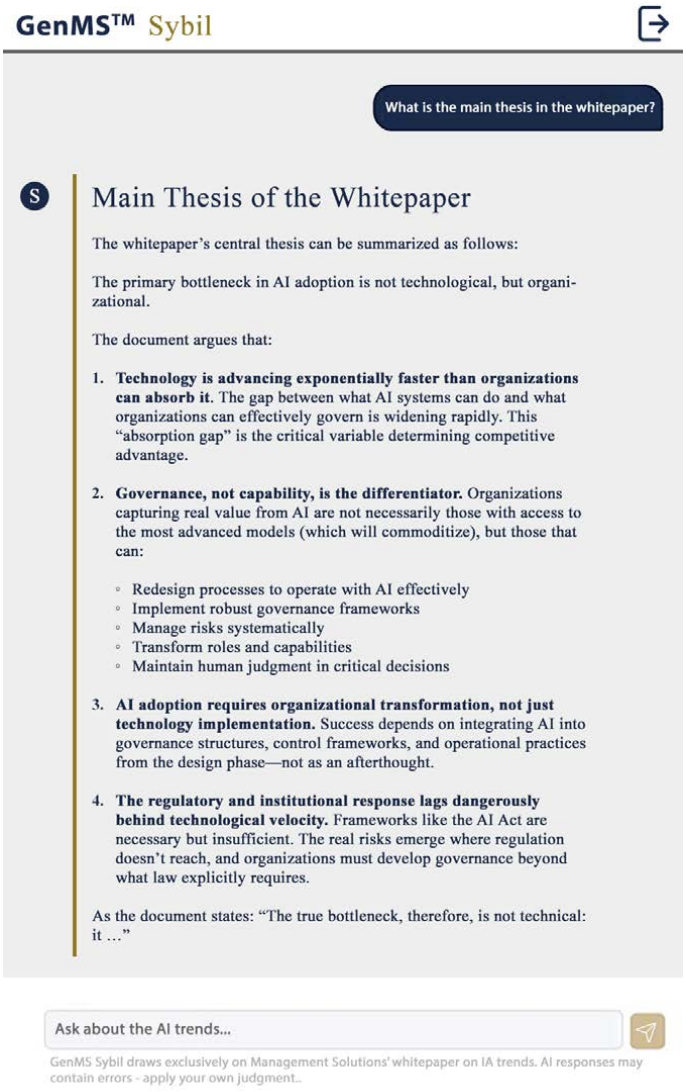
The GenMS™ Sybil system prompt is several pages long. It codifies the behavioral guardrails, the operational limits, the handling of out-of-scope questions and the ethical principles governing system responses. Its contents are not published in full for security reasons. Its length reflects a principle that this document articulates in the ethics section: the gap between the stated values and the actual behavior of an AI system is closed in the specific instructions that govern it, not in the principles that frame it.

Trends in action

GenMS™ Sybil does not illustrate the trends this paper discusses: it implements them. The democratization of generative AI made it possible for profiles without software engineering specialization to produce a production system. Vibe coding was the construction method, not the metaphor. LLMOps structured a process that would otherwise have been unrepeatable. The profiles involved combine business knowledge with the ability to run cognitive systems: the profile that talent analysis identifies as the scarcest and most decisive. AI audited AI at the security phase. GenMS Atlas applied systematic validation where ad hoc validation would have been insufficient. Regulation was a design criterion, not a requirement for subsequent compliance.

The bottleneck that this paper repeatedly identifies as the limiting factor in actual AI adoption is not technological. This case does not show that it is easy to overcome; it shows that it is surmountable, and that the distance between describing it and crossing it is, today, fundamentally a matter of organizational readiness.

Fig. 11. GenMS™ Sybil.



08 | Conclusion

"The question is no longer whether AI will transform our world. The question is whether we will govern this transformation together or let it govern us".

António Guterres[229]

The trends discussed converge on one central variable: the speed at which the gap between what AI systems can do and what organizations can govern is accelerating. That gap - not the language model, the autonomous agent or the robot - is the central management object of the coming years.

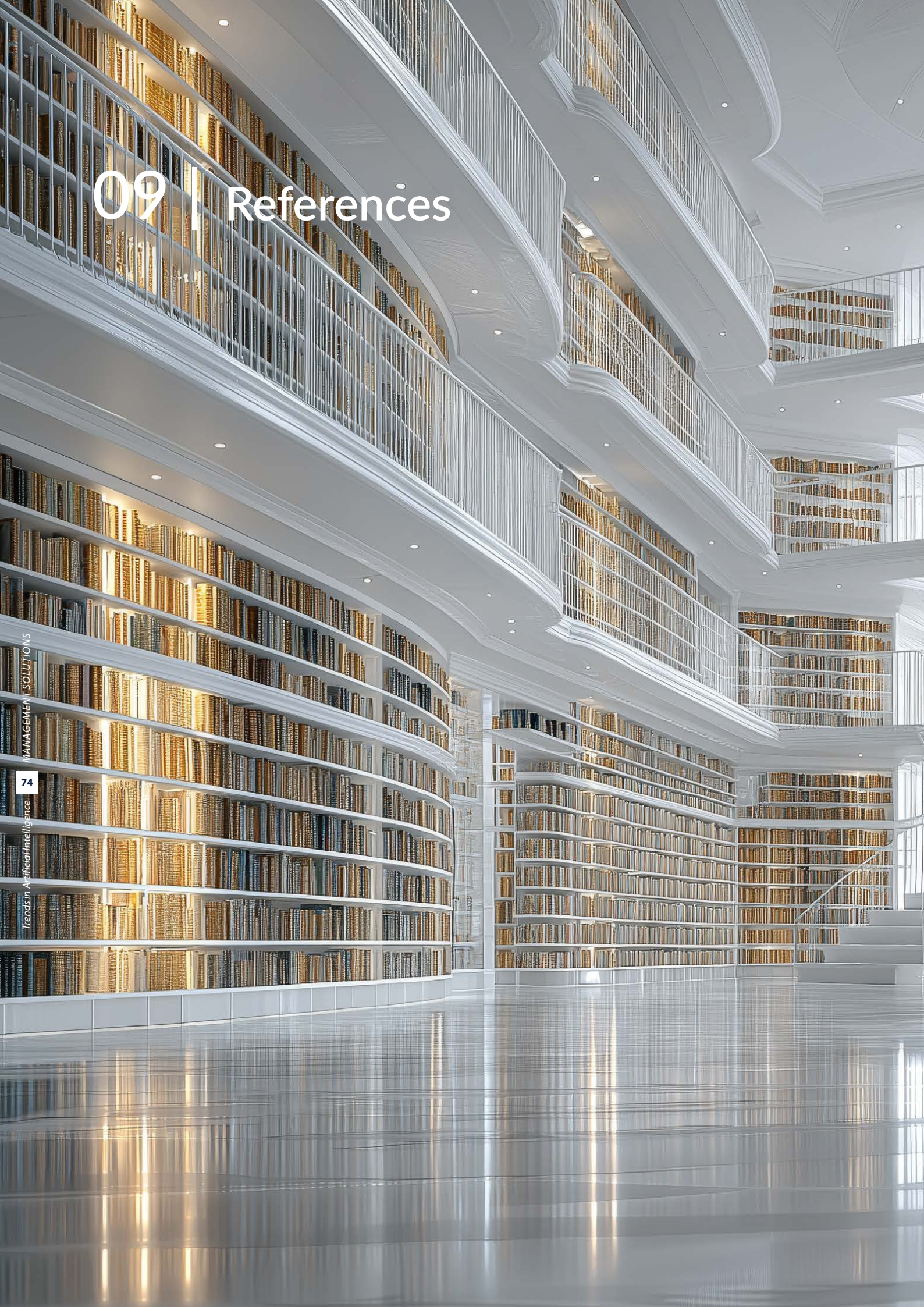
The sustainable competitive differential lies not in access to the most advanced models, which will become progressively commoditized, but in the speed and rigor with which an organization is able to redesign itself to operate with them effectively and responsibly. Organizations that are capturing real value share one characteristic: they have understood AI adoption as an organizational transformation, not a technology project. Governance that enables rather than slows down, training that transforms rather than certifies, risk frameworks that manage uncertainty without sacrificing speed.

Regulatory frameworks, technical standards and ethical principles are a necessary but not sufficient condition. The AI Act classifies risk; ISO and NIST standards structure management; ethics frameworks produce operating principles. None of these alone resolves the question that underlies several of the trends analyzed: what kind of entity is being deployed, what relationship it establishes with the people who use it, and what obligations does that generate beyond what current regulation explicitly requires. It is precisely where regulation does not reach that the risks have the least visibility and the greatest potential for harm.

The gap between the technological capability curve and the organizational absorption curve is widening every day. Technology advances independently of the internal decision-making speed of each organization; organizational adjustment, on the other hand, depends on it. This asymmetry is what makes AI governance a strategic variable of the first order, comparable in impact to technical capacity itself.

The stakes transcend individual competitiveness. The distribution of the benefits of AI, the preservation of human judgment in decisions that require it, the ability of institutions to maintain legitimacy in systems that evolve faster than the structures designed to govern them: these are dimensions that no organization can manage in isolation, and for which the institutional response is, at the moment, significantly behind the speed of the problem.

227 António Guterres (b. 1949), ninth secretary-general of the United Nations and former prime minister of Portugal, who has been a prominent driver of the international agenda on AI governance and emerging technology risks.



09 | References

ABILab (2026). AI Banking (R)evolution: oltre la scelta. Rapporto AI Hub.

AESIA (2026). Agencia Española de Supervisión de Inteligencia Artificial. https://aesia.digital.gob.es/es

AI Act (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

AI Board (2026). Governance and coordination; AI board meetings. https://digital-strategy.ec.europa.eu/en/policies/ai-board

AICerts (2026). Generative AI Phishing Boosts Clicks, Reshapes Cyber Risk. https://www.aicerts.ai/news/generative-ai-phishing-boosts-clicks-reshapes-cyber-risk/

Altman (2025a). Three Observations. https://blog.samaltman.com/three-observations

Altman (2025b). The Gentle Singularity. https://blog.samaltman.com/the-gentle-singularity

Altman (2024a). Could AI create a one-person unicorn? Fortune. https://finance.yahoo.com/news/could-ai-create-one-person-120000722.html

Altman (2024b). The Intelligence Age. Blog personal. https://ia.samaltman.com/

Amazon (2023). Amazon announces 8 innovations to better deliver for customers, support employees, and give back to communities around the world. https://www.aboutamazon.com/news/operations/amazon-delivering-the-future-2023-announcements

Amodei (2024a). Machines of loving grace. https://www.darioamodei.com/essay/machines-of-loving-grace

Amodei (2024b). Machines of Loving Grace: How AI Could Transform the World for the Better. Blog personal. https://www.darioamodei.com/essay/machines-of-loving-grace

Amodei (2025). Technology in the World, Annual Meeting Davos 2025, World Economic Forum. https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2025/sessions/technology-in-the-world/

Amodei (2026). The Adolescence of Technology. https://www.darioamodei.com/essay/the-adolescence-of-technology

Anthropic (2025). Anthropic Economic Index – September 2025 Report. https://www.anthropic.com/research/anthropic-economic-index-september-2025-report

Anthropic (2026). Claude's new constitution. Anthropic. https://www.anthropic.com/news/claude-new-constitution

Australia (2025). Australia's AI Ethics Principles. https://www.industry.gov.au/publications/australias-ai-ethics-principles

Backlinko (2025). ChatGPT / OpenAI Statistics: How Many People Use ChatGPT? https://backlinko.com/chatgpt-stats

Baker McKenzie (2025). Navigating Labor's Response to AI: Proactive Strategies for Multinational Employers Across the Atlantic. https://www.theemployerreport.com/2025/06/navigating-labors-response-to-ai-proactive-strategies-for-multinational-employers-across-the-atlantic/

Barkhuus (2003). Is Context-Aware Computing Taking Control Away from the User? Three Levels of Interactivity Examined. UbiComp 2003. Springer. https://doi.org/10.1007/978-3-540-39653-6_12

Batty (2024). Digital Twins in City Planning. *Nature Computational Science*, 4, 192–199. https://doi.org/10.1038/s43588-024-00606-5

Bettencourt (2024). Recent Achievements and Conceptual Challenges for Urban Digital Twins. Nature Computational Science, 4, 150–153. https://doi.org/10.1038/s43588-024-00604-7

Bimpas (2024). Leveraging Pervasive Computing for Ambient Intelligence: A Survey on Recent Advancements, Applications and Open Challenges. Computer Networks, 239, 110156. https://doi.org/10.1016/j.comnet.2023.110156

Bletchley Declaration (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Bloomberg (2026). AI Startup Nabs $100 Million to Help Firms Predict Human Behavior. https://www.bloomberg.com/news/articles/2026-02-12/ai-startup-nabs-100-million-to-help-firms-predict-human-behavior

Boston Dynamics (2025a). An Electric New Era for Atlas. https://bostondynamics.com/blog/electric-new-era-for-atlas/

Boston Dynamics (2025b). Large Behavior Models and Atlas Find New Footing. https://bostondynamics.com/blog/large-behavior-models-atlas-find-new-footing/

Brown (2025). AI's War in the Courtroom: Copyright Disputes Spike in 2025. https://www.bestlawfirms.com/articles/ai-war-in-the-courtroom-copyright-disputes-spike-in-2025/7186

Business Insider (2025a). Walmart just showed off its new AI-powered warehouses — take a look inside. https://www.businessinsider.com/see-inside-walmart-high-tech-refrigerated-grocery-warehouse-2024-7

Business Insider (2025b). The guy who coined 'vibe coding' predicts it will 'terraform software and alter job descriptions'. https://www.businessinsider.com/andrej-karpathy-coined-vibecoding-ai-prediction-2025-12

Cambridge (2025). Navigating China's regulatory approach to generative artificial intelligence and large language models. https://www.cambridge.org/core/journals/cambridge-forum-on-ai-law-and-governance/article/navigating-chinas-regulatory-approach-to-generative-artificial-intelligence-and-large-language-models/969B2055997BF42DE693B7A1A1B4E8BA

Centre for European Policy (2026). Competition in Generative AI: Updated Assessment. cepInput No. 1/2026. https://www.cep.eu/fileadmin/user_upload/cep.eu/Studien/cepInput_Competition_in_Generative_AI/cepInput_Competition_in_GenAI_Updated_Assessment.pdf

Chatgptiseatingtheworld (2026). Updated Master chart of copyright, DMCA and other claims in suits v. AI (Dec. 5, 2025). https://chatgptiseatingtheworld.com/2025/12/03/updated-master-chart-of-copyright-dmca-and-other-claims-in-suits-v-ai-dec-3-2025/

Chen (2025). Need Help? Designing Proactive AI Assistants for Programming. CHI 2025. ACM. https://doi.org/10.1145/3706598.3714002

Cheong (2025). E2E Process Automation Leveraging Generative AI and IDP-Based Automation Agent: A Case Study on Corporate Expense Processing. https://arxiv.org/abs/2505.20733

Christensen (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press.

CKGSB (2025). Cheung Kong Graduate School of Business. Banking on data: How MYbank is revolutionizing supply chain finance. CKGSB Knowledge. https://english.ckgsb.edu.cn/knowledge/article/unleashing-innovation-in-china-series-banking-on-data-how-mybank-is-revolutionizing-supply-chain-finance/

Corrêa (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Covington (2025). New Artificial Intelligence Legislation in Mexico. Global Policy Watch. https://www.globalpolicywatch.com/2025/03/new-artificial-intelligence-legislation-in-mexico/

CrowdStrike (2025). CrowdStrike Advances Next-Gen SIEM with Threat Hunting Across Data Sources, AI-Driven UEBA. https://www.crowdstrike.com/en-us/blog/crowdstrike-advances-next-gen-siem-capabilities/

Cyberhaven (2025). AI Adoption and Risk Report Q2 2025. https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Labs%20-%202025%20AI%20Adoption%20&%20Risk%20Report.pdf

Darktrace (2025). New Report Finds that 78% of Chief Information Security Officers Globally are Seeing a Significant Impact from AI-Powered Cyber Threats – up 5% from last year. https://www.darktrace.com/news/new-report-finds-that-78-of-chief-information-security-officers-globally-are-seeing-a-significant-impact-from-ai-powered-cyber-threats

DBS Bank (2024). DBS AI-Powered Digital Transformation. https://www.dbs.com/artificial-intelligence-machine-learning/artificial-intelligence/dbs-ai-powered-digital-transformation.html

Dealroom (2025). AI startups: Revenue per employee benchmarks. https://x.com/dealroomco/status/1914264599505018989

Deutsche Bank (2025). Claudio de Sanctis, Head of Private Bank, Deutsche Bank AG Private Bank. Investor Deep Dive 2025. https://investor-relations.db.com/files/documents/other-presentations-and-events/2025/IDD-2025-Script-Private-Bank-Claudio-de-Sanctis.pdf

DHL (2024). DHL Supply Chain Passes Unprecedented 500 Million Picks Milestone Using Locus Robotics Autonomous Mobile Robots. https://www.dhl.com/es-en/home/press/press-archive/2024/dhl-supply-chain-passes-unprecedented-500-million-picks-milestone-using-locus-robotics-autonomous-mobile-robots.html

EBA (2021). EBA Discussion Paper on Machine Learning for IRB Models. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf

ECB (2025). ECB Guide to Internal Models. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf

EDPB (2025). AI Privacy Risks & Mitigations Large Language Models (LLMs). https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en

Epoch (2025a). AI Benchmarking. https://epoch.ai/benchmarks

Epoch (2025b). How much power will frontier AI training demand in 2030? https://epoch.ai/blog/power-demands-of-frontier-ai-training

ESOMAR (2024). Global Market Research 2024. https://shop.esomar.org/knowledge-center/library?publication=3019

Eurostat (2025). 32.7% of EU people used generative AI tools in 2025. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251216-3

Financial News London (2025). Deutsche Bank to roll out 'banking butlers' for affluent clients. https://www.fnlondon.com/articles/deutsche-bank-to-roll-out-banking-butlers-for-ultra-wealthy-clients-77e0349a

Figure (2025). F.02 Contributed to the Production of 30,000 Cars at BMW. https://www.figure.ai/news/production-at-bmw

FirstPageSage (2025). ChatGPT Usage Statistics: December 2025. https://firstpagesage.com/seo-blog/chatgpt-usage-statistics/

Fortune (2025a). Deloitte allegedly cited AI-generated research in a million-dollar report for a Canadian provincial government. https://fortune.com/2025/11/25/deloitte-caught-fabricated-ai-generated-research-million-dollar-report-canada-government/

Fortune (2025b). Elon Musk reveals massive plans for Tesla and Optimus—'Things are really going to go ballistic next year'. https://fortune.com/2025/01/30/elon-musk-reveals-massive-plans-tesla-optimus-self-driving-cars-humanoid-robots/

Fortune (2025c). AI enabled Klarna to halve its workforce—now, the CEO is warning other tech leaders to be honest about the risks. https://fortune.com/2025/10/10/klarna-ceo-sebastian-siemiatkowski-halved-workforce-says-tech-ceos-sugarcoating-ai-impact-on-jobs-mass-unemployment-warning/

Gartner (2025a). Hype Cycle for Artificial Intelligence. https://www.gartner.com/en/newsroom/press-releases/2025-08-05-gartner-hype-cycle-identifies-top-ai-innovations-in-2025

Gartner (2025b). Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

Google (2023). Practitioners Guide to MLOps: A framework for continuous delivery and automation of machine learning. https://cloud.google.com/resources/mlops-whitepaper

Google DeepMind (2025). AlphaFold: Science and impact. https://deepmind.google/science/alphafold/

Google Quantum AI (2024). Quantum error correction below the surface code threshold. *Nature*, 638, 920–926. https://doi.org/10.1038/s41586-024-08449-y

Grieves-Vickers (2017). Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. En Kahlen, F.J. et al. (eds.), *Transdisciplinary Perspectives on Complex Systems*. Springer. https://doi.org/10.1007/978-3-319-38756-7_4

Gu (2025). Large Language Models for Constructing and Optimizing Machine Learning Workflows: A Survey. https://arxiv.org/html/2411.10478v1

Heydari (2025). Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions. Sensors, 25(10), 3191. https://doi.org/10.3390/s25103191

HelpNetSecurity (2025). 67% of daily security alerts overwhelm SOC analysts. https://www.helpnetsecurity.com/2023/07/20/soc-analysts-tools-effectiveness/

Hernández-Torres (2025). Challenges and Opportunities of Ambient Intelligence (AmI) in the 21st Century: A Historical Review. Evolutionary Intelligence, 18, 80. https://doi.org/10.1007/s12065-025-01010-z

Huang (2021). Power of data in quantum machine learning. *Nature Communications*, 12, 2631. https://doi.org/10.1038/s41467-021-22539-9

Hyundai (2025). Hyundai Motor Group to Unveil AI Robotics Strategy at CES 2026. https://www.hyundai.com/worldwide/en/newsroom/detail/0000001093

IBM (2025). Cost of a Data Breach Report 2025. https://www.ibm.com/reports/data-breach

IBM (2025b). Chief AI Officers cut through complexity to create new paths to value. https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/chief-ai-officer

Index Ventures (2026). Life, the Universe, and Simile: Leading Simile's $100M Series A. https://www.indexventures.com/perspectives/life-the-universe-and-simile-leading-similes-series-a/

iDanae (3Q20). Cátedra iDanae (UPM–Management Solutions). MLOps, a key element in the digital ecosystem. 2020. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2020/10/Idanae-3Q20.pdf

iDanae (2Q23). Cátedra iDanae (UPM–Management Solutions). Large Language Models: a new era for Artificial Intelligence. 2023. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2023/07/Idanae-2Q23.pdf**iDanae (1Q24).** Cátedra iDanae (UPM–Management Solutions). Towards a sustainable Artificial Intelligence. 2024. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2024/04/Idanae-1Q24-VDef.pdf

iDanae (1Q25). Cátedra iDanae (UPM–Management Solutions). The challenge of biases in the construction of Artificial Intelligence systems. 2025. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2025/04/Idanae-1Q25.pdf

iDanae (2Q25). Cátedra iDanae (UPM–Management Solutions). GenAI: an approach to multi-agents systems. 2025. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2025/07/Idanae-2Q25.pdf

IEA (2025a). Electricity 2025: Analysis and forecast to 2027. International Energy Agency. https://www.iea.org/reports/electricity-2025

IEA (2025b). World Energy Outlook Special Report on Energy and AI. International Energy Agency. https://www.iea.org/reports/energy-and-ai

IMD (2023). In the Field with Ping An. *IMD Business School*. https://www.imd.org/research-knowledge/digital/articles/in-the-field-with-ping-an/

IMF (2024). Gen-AI: Artificial Intelligence and the Future of Work. https://www.imf.org/-/media/files/publications/sdn/2024/english/sdnea2024001.pdf

ILO (2025a). International Labour Organization. Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality. https://www.ilo.org/sites/default/files/2025-05/WP140_web.pdf

ILO (2025b). International Labour Organization. Governing AI in the World of Work: A review of global ethics guidelines. https://www.ilo.org/resource/article/governing-ai-world-work-review-global-ethics-guidelines

ILO (2025c). International Labour Organization. Global Case Studies of Social Dialogue on AI and Algorithmic Management. https://www.ilo.org/sites/default/files/2025-07/wp144_web.pdf

Inter-Parliamentary Union (2025). Parliamentary actions on AI policy. https://www.ipu.org/impact/democracy-and-strong-parliaments/artificial-intelligence/parliamentary-actions-ai-policy

Ironscales (2025). Ironscales Fall 2025 Threat Report. https://ironscales.com/hubfs/Landing%20Page%20Assets/Fall%202025%20Threat%20Report/2025%20Fall%20Threat%20Report_Beyond%20Detection_Reality%20of%20Deepfake%20Attacks%202.pdf

ISO/IEC (2023). ISO/IEC 42001:2023 - Artificial Intelligence Management Systems. https://www.iso.org/standard/42001

Jones (2025). Large Language Models Pass the Turing Test. https://arxiv.org/abs/2503.23674

Jumpcloud (2025). How Effective Is AI for Cybersecurity Teams? 2025 Statistics. https://jumpcloud.com/blog/how-effective-is-ai-for-cybersecurity-teams

KnowBe4 (2025). Phishing Threat Trends Report. https://www.knowbe4.com/hubfs/Phishing-Threat-Trends-2025_Report.pdf

Krijger, J. (2023). Operationalising ethics for AI in the financial industry: Insights from the Volksbank case study. *Journal of Digital Banking*, 8(3), 220–241. https://doi.org/10.69554/YQZC2796

Kumar (2020). Adversarial Machine Learning - A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2023. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

Kusumegi et al. (2025). Scientific production in the era of large language models. Science. https://www.science.org/doi/10.1126/science.adw3000

Li (2025). Pitfalls and prospects of quantum machine learning. *Nature Computational Science*, 5, 1095–1097. https://doi.org/10.1038/s43588-025-00914-6

Liu (2024). Towards provably efficient quantum algorithms for large-scale machine-learning models. *Nature Communications*, 15, 434. https://doi.org/10.1038/s41467-023-43957-x

Lukac (2025). Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI. https://doi.org/10.1056/AIoa2501000

Management Solutions (2023). Explainable Artificial Intelligence (XAI): Challenges of model interpretability. 2023. https://www.managementsolutions.com/en/microsites/whitepapers/explainable-artificial-intelligence

Management Solutions (3Q23). Follow-up report on Machine Learning for IRB models. Technical note on regulations, 3Q23, 2023.

https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/follow-report-machine-learning-irb-models

Management Solutions (4Q24). Artificial Intelligence Act. Technical note on regulations, 4Q24, 2024. https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/proposal-regulation-european-approach-artificial-intelligence

Management Solutions (4Q24a). Artificial Intelligence: regulatory landscape. Technical note on regulations, 4Q24, 2024. https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/artificial-intelligence-regulatory-landscape

Marwala (2025). Why the Need for Governing Ambient Intelligence Has Never Been More Urgent. United Nations University. https://unu.edu/article/why-need-governing-ambient-intelligence-has-never-been-more-urgent

Mascelli (2025). Harvest Now Decrypt Later: Examining Post-Quantum Cryptography and the Data Privacy Risks for Distributed Ledger Networks. *Finance and Economics Discussion Series* 2025-093. Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2025.093

Microsoft (2026). AI-powered SIEM, built for modern security. https://marketingassets.microsoft.com/gdc/gdccKuCLQ/original

MIT Technology Review (2024). What's next for AlphaFold: A conversation with a Google DeepMind Nobel laureate. https://www.technologyreview.com/2025/11/24/1128322/whats-next-for-alphafold-a-conversation-with-a-google-deepmind-nobel-laureate/

MITRE (2025). ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems). https://atlas.mitre.org/

Moroni (2025). Insurmountable Limitations of City-Scale Digital Twins? On Urban Knowledge and Planning. Computational Urban Science. Springer Nature. https://doi.org/10.1007/s43762-025-00174-0

Nadella (2025). LinkedIn Post. https://www.linkedin.com/posts/satyanadella_just-wrapped-our-earnings-call-and-wanted-activity-7389433821181562880-GnMz/

Nissenbaum (1996). Accountability in a Computerized Society. Science and Engineering Ethics, 2, 25–42. https://link.springer.com/article/10.1007/BF02639315

NIST (2023). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework

NIST (2024a). Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. https://www.nist.gov/publications/secure-software-development-practices-generative-ai-and-dual-use-foundation-models-ssdf

NIST (2024b). Post-Quantum Cryptography Standards: FIPS 203, FIPS 204, FIPS 205. National Institute of Standards and Technology. https://csrc.nist.gov/projects/post-quantum-cryptography

Notateslaapp (2025). Tesla Eyes $20K Price Target For Optimus, Extremely Fast Production Ramp. https://www.notateslaapp.com/news/3314/tesla-eyes-20k-price-target-for-optimus-extremely-fast-production-ramp

OECD (2024). A Sectoral Taxonomy of AI Intensity. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/12/a-sectoral-taxonomy-of-ai-intensity_c2baae71/1f6377b5-en.pdf

OECD (2025a). Emerging Divides in the Transition to Artificial Intelligence. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/emerging-divides-in-the-transition-to-artificial-intelligence_eeb5e120/7376c776-en.pdf

OECD (2025b). The effects of generative AI on productivity, innovation and entrepreneurship. OECD Artificial Intelligence Papers, no. 39. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/the-effects-of-generative-ai-on-productivity-innovation-and-entrepreneurship_da1d085d/b21df222-en.pdf

OECD (2026). AI Use by Individuals Surges Across the OECD as Adoption by Firms Continues to Expand. https://www.oecd.org/en/about/news/announcements/2026/01/ai-use-by-individuals-surges-across-the-oecd-as-adoption-by-firms-continues-to-expand.html

Ouyang (2025). FELA: A Multi-Agent Evolutionary System for Feature Engineering of Industrial Event Log Data. https://arxiv.org/html/2510.25223

OWASP (2025). OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Oxford (2025). AI Regulation: The Politics of Fragmentation and Regulatory Capture. https://blogs.law.ox.ac.uk/oblb/blog-post/2025/06/ai-regulation-politics-fragmentation-and-regulatory-capture

Parfit (1984). Reasons and Persons. Oxford University Press.

Park (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM. https://doi.org/10.1145/3586183.3606763

Park (2024). Generative Agent Simulations of 1,000 People. https://arxiv.org/abs/2411.10109

Patel (2025). U.S. AI Law & Policy Explained. Enterprise AI Governance. https://oliverpatel.substack.com/p/us-ai-law-and-policy-explained

Pew (2025). Pew Research Institute. How the U.S. Public and AI Experts View Artificial Intelligence. https://www.pewresearch.org/

Phishcare (2025). Top 10 Deepfake Phishing Scams. https://phishcare.com/top-10-deepfake-phishing-scams/

Pixiebrix (2025). Top Chief AI Officers of 2025. https://www.pixiebrix.com/reports/top-ai-officers-of-2025

PR Newswire (2023). SlashNext's 2023 State of Phishing Report Reveals a 1,265% Increase in Phishing Emails Since the Launch of ChatGPT in November 2022, Signaling a New Era of Cybercrime Fueled by Generative AI. https://www.prnewswire.com/news-releases/slashnexts-2023-state-of-phishing-report-reveals-a-1-265-increase-in-phishing-emails-since-the-launch-of-chatgpt-in-november-2022--signaling-a-new-era-of-cybercrime-fueled-by-generative-ai-301971557.html

Proofpoint (2025). AI Threat Detection. https://www.proofpoint.com/us/threat-reference/ai-threat-detection

Pu (2025). Assistance or Disruption? Exploring and Evaluating the Design and Trade-offs of Proactive AI Programming Support. CHI 2025. ACM. https://doi.org/10.1145/3706598.3713357

Quartz (2025). AI powers smaller startups toward a new era of unicorns. https://qz.com/unicorn-entrepreneur-founder-solo-ai-startup-automation-workforce

Rawls (1971). A Theory of Justice. Harvard University Press.

Ryt Bank (2025). The World's First AI-Powered Bank. https://www.rytbank.my/

Sacra (2026). Cursor revenue, valuation & funding. https://sacra.com/c/cursor/

Shan (2024). Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models. https://doi.org/10.3390/info16020087

Shankar (2021). Towards Observability for Machine Learning Pipelines. DOI:10.48550/arXiv.2108.13557. (PDF) Towards Observability for Machine Learning Pipelines

SoSafe (2025). Global businesses face escalating AI risk, as 87% hit by AI cyberattacks. https://sosafe-awareness.com/company/press/global-businesses-face-escalating-ai-risk-as-87-hit-by-ai-cyberattacks/

SQ Magazine (2025). AI Cyber Attacks Statistics 2025: How Attacks, Deepfakes & Ransomware Have Escalated. https://sqmagazine.co.uk/ai-cyber-attacks-statistics/

Stanford (2023a). Stanford Encyclopedia of Philosophy. Ethics of Artificial Intelligence and Robotics. https://plato.stanford.edu/entries/ethics-ai/

Stanford (2023b). Stanford Encyclopedia of Philosophy. Philosophy of Artificial Intelligence. https://plato.stanford.edu/entries/artificial-intelligence/

Stanford (2025). The 2025 AI Index Report. https://hai.stanford.edu/ai-index/2025-ai-index-report

Stone (2025). Navigating MLOps: Insights into Maturity, Lifecycle, Tools, and Careers. https://arxiv.org/html/2503.15577v1

Tesla Car World (2025). Elon Musk Unveils Tesla Bot Gen 3 Real Homemaker Updates. https://www.youtube.com/watch?v=HR1HrrneNHs&t=1s

Teslarati (2025). Tesla Optimus' pilot line will already have an incredible annual output. https://www.teslarati.com/tesla-optimus-pilot-line-will-already-have-an-incredible-annual-output/

The Network Installers (2025). AI Cyber Threat Statistics. https://thenetworkinstallers.com/es/blog/ai-cyber-threat-statistics/

Thompson (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. American Political Science Review, 74(4), 905–916. https://www.cambridge.org/core/journals/american-political-science-review/article/abs/moral-responsibility-of-public-officials-the-problem-of-many-hands/39DD3FAB7BF7DC7A242407143674F22B

Toyota Research Institute (2025). AI-Powered Robot by Boston Dynamics and Toyota Research Institute Takes a Key Step Towards General-Purpose Humanoids. https://www.tri.global/news/ai-powered-robot-boston-dynamics-and-toyota-research-institute-takes-key-step-towards-general

UK AI Safety Institute (2026). International AI Safety Report 2026. UK Government. https://www.gov.uk/government/publications/international-ai-safety-report-2026

UK DSIT (2023). UK Department for Science, Innovation and Technology. Public Attitudes to Data and AI: Tracker Survey (Report). https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey

UK Government (2023). A pro-innovation approach to AI regulation. Policy paper. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

UNDP (2025). Human Development Report 2025. United Nations Development Programme. https://hdr.undp.org/content/human-development-report-2025

UNESCO (2023). Guidance for Generative AI in Education and Research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

United Nations (2025). The Sustainable Development Goals Report 2025. United Nations, Department of Economic and Social Affairs. https://unstats.un.org/sdgs/report/2025/

WatchGuard (2025). Evasive Malware Surges 40% in WatchGuard's Latest Internet Security Report. https://www.watchguard.com/es/wgrd-news/blog/evasive-malware-surges-40-watchguards-latest-internet-security-report

WIPO (2025). WIPO Conversation on Intellectual Property and Frontier Technologies. https://www.wipo.int/en/web/frontier-technologies/frontier_conversation

World Bank (2024). Global Trends in AI Governance. https://documents1.worldbank.org/curated/en/099120224205026271/pdf/P1786161ad76ca0ae1ba3b1558ca4ff88ba.pdf

World Bank (2025). World Development Report 2025: Standards for Development. World Bank. https://www.worldbank.org/en/publication/wdr2025

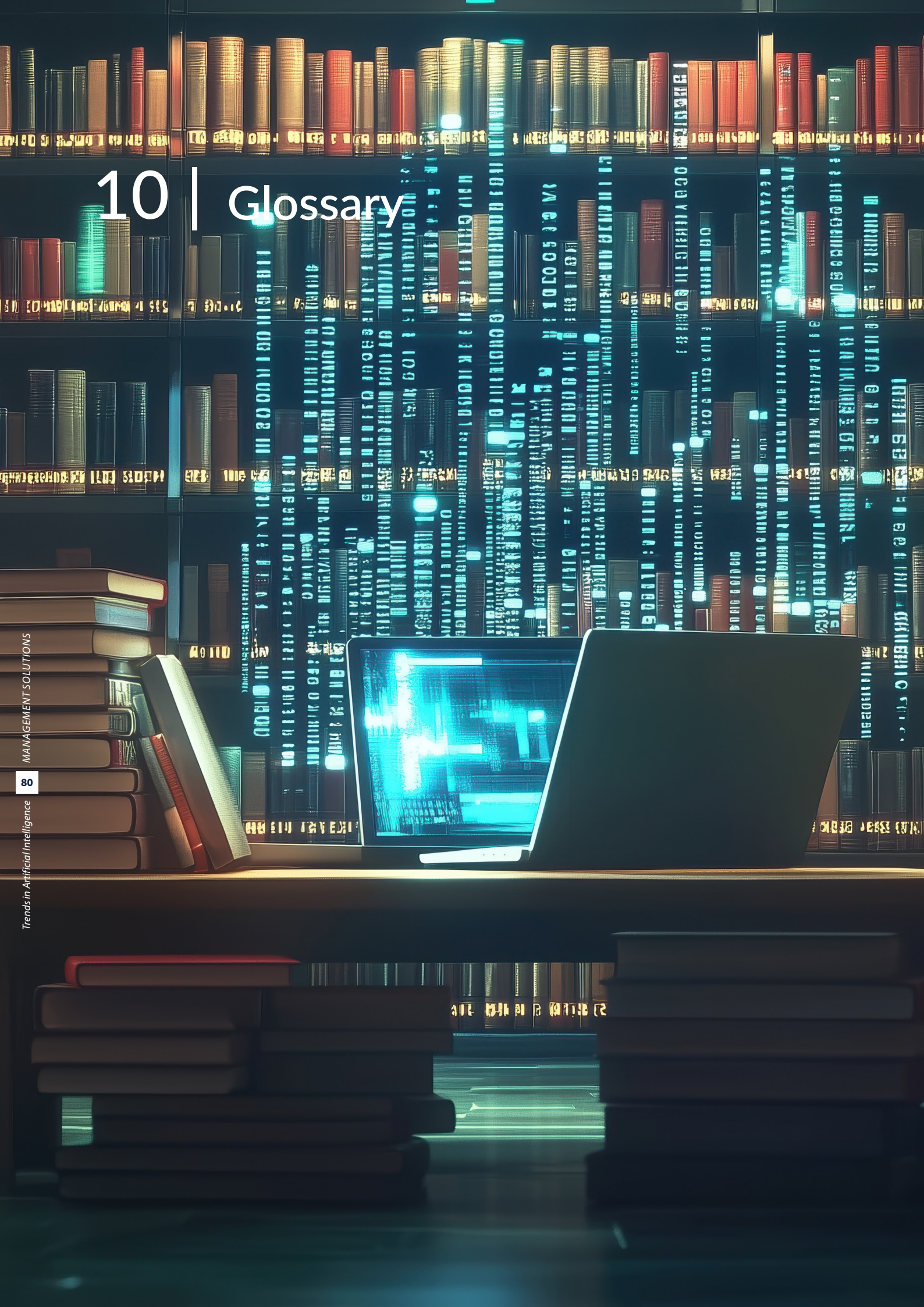
World Economic Forum (2025a). AI in Action: Beyond Experimentation to Transform Industry. https://reports.weforum.org/docs/WEF_AI_in_Action_Beyond_Experimentation_to_Transform_Industry_2025.pdf

World Economic Forum (2025b). Transforming Consumer Industries in the Age of AI. https://reports.weforum.org/docs/WEF_Transforming_Consumer_Industries_in_the_Age_of_AI_2025.pdf

World Health Organization (2024). Ethics and Governance of Artificial Intelligence for Health. https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content

Yuan et al. (2025). The Impact of AI Adoption in the Workplace on Employees: A Systematic Review. https://www.researchgate.net/publication/396219396_The_Impact_of_AI_Adoption_in_the_Workplace_on_Employees_A_Systematic_Review

10 | Glossary



Absorption gap: Increasing distance between the technological capability curve of AI (exponential, self-accelerating) and the organizational absorption curve (process redesign, role re-engineering, governance).

Adversarial evasion: Attack that introduces imperceptible perturbations in the inputs of a model to cause incorrect classifications or responses during inference, without modifying the model itself.

Agentic AI: AI systems that plan, execute complex tasks and operate autonomously on real corporate infrastructures, beyond responding to prompts. Incremental capabilities: persistent state, dynamic planning, execution on real systems and multi-agent orchestration.

AGI (Artificial General Intelligence): AI system capable of performing any cognitive task that can be performed by a human being, with transferable generalization across domains. A debated strategic horizon whose precise definition lacks scientific consensus.

AI Act: Regulation (EU) 2024/1689, first comprehensive legal framework on AI. Classifies systems by risk level (unacceptable, high, limited, minimal) and imposes structural obligations on high-risk systems. Penalties of up to €35 million or 7% of global turnover.

AI Board: Forum for coordination and common interpretation between the European Commission and the national AI supervisory authorities, created by the AI Act.

AI Office: Central technical body of the European Commission responsible for the supervision of general purpose AI models (GPAI) under the AI Act.

AI-enhanced: Organizational model in which AI is used to optimize existing processes without redesigning them from AI capabilities. Predominant stage in most organizations today.

AI-first: Organizational model in which processes and structure are designed based on AI capabilities, assigning to human judgment only those tasks where its comparative advantage is unequivocal.

AI-only: Hypothetical organizational model in which core operations functionally dispense with human labor.

AIMS (AI Management System): AI management system conforming to ISO/IEC 42001, equivalent to ISO 27001 for cybersecurity but specific to AI: covers policies, impact assessments, vendor control and continuous monitoring.

Ambient AI: AI that operates without being explicitly invoked: it continuously observes the context, infers needs and acts proactively. The interface disappears; the environment itself becomes the point of interaction.

Ambient scribe: Ambient AI system deployed in clinical settings that listens to the doctor-patient conversation and automatically generates documentation of the encounter without explicit instruction from the user.

BEC (Business Email Compromise): A type of cyberattack in which the identity of an executive or supplier is impersonated to divert funds or extract information. Generative AI powers it through highly credible audio and video deepfakes.

Brussels effect: Phenomenon whereby EU regulation (AI Act, GDPR) forces global companies to adapt their products to European standards due to market volume, de facto exporting this regulation to the rest of the world.

CAIO / CDAIO (Chief AI Officer / Chief Data and AI Officer): Executive function responsible for the strategic leadership of AI in an organization.

Citizen data scientist: Non-technical professional capable of performing basic data analysis with visual tools. Generative AI goes beyond this concept by delivering advanced analytical capabilities directly to non-technical end users.

Cryptography resistant to quantum attacks (PQC): Set of cryptographic algorithms designed to resist attacks by quantum computers. NIST published the first standards in 2024 (FIPS 203, 204, 205).

Dark LLM: Language models modified specifically for cybercrime (WormGPT, FraudGPT, GhostGPT). They generate malware, exploits and social engineering campaigns without ethical restrictions. Marketed on the dark web with technical support.

Data poisoning: Adversarial attack that consists of injecting malicious data into the training set of a model to degrade its behavior or introduce biases controlled by the attacker.

Deepfake: AI-generated synthetic audiovisual content that impersonates the appearance or voice of a real person. Used in BEC, fraud and disinformation attacks with significantly higher success rate than traditional phishing.

Differential privacy: Technique that adds controlled statistical noise to data or results to prevent identification of individuals, preserving aggregate statistical usefulness.

Digital Twin: Dynamic simulation of a physical or human system that is updated in real time with data from the real system. It works with high reliability on deterministic physical systems; LLMs have opened their extension to human behavior and complex social systems.

DPIA (Data Protection Impact Assessment): Data protection impact assessment required by GDPR (Art. 35). The EDPB considers it mandatory in most LLM deployments given their systemic processing of personal data.

Edge AI / TinyML: Ability to run AI models directly on devices such as smartphones, wearables and sensors without relying on cloud connectivity. Key enabler of Ambient AI.

Ethics washing: Phenomenon whereby an organization publishes ethical AI principles without translating them into operational controls, concrete responsibilities or auditing mechanisms.

Explainability: Ability to describe the inner workings of an AI model in a way that is understandable to different audiences (regulator, customer, employee). Technical requirement in regulated models; implementable in ML with techniques such as SHAP, LIME or sensitivity analysis.

Feature engineering: Process of building predictive variables from raw data to feed ML models. One of the most expert knowledge intensive phases of the classical ML lifecycle; significantly accelerated by generative AI.

Federated learning: Distributed training paradigm in which data remains on local devices and only model updates are shared. Mitigates privacy risks in LLMs at the cost of increased operational complexity.

GDPR (General Data Protection Regulation): Data Protection Regulation (EU) 2016/679. Its principles of minimization, right to be forgotten and transparency present structural tensions with the architecture and lifecycle of LLMs.

Generative AI: Family of models capable of generating original content (text, images, audio, video, code) in response to natural language instructions. Based on large-scale transformer architectures.

GPAI (General Purpose AI): General purpose AI model (e.g., GPT, Claude, Gemini) capable of performing a wide variety of tasks.

Gradient boosting: ML technique that combines multiple decision trees sequentially to improve prediction.

Hallucination: Intrinsic behavior of generative models whereby they produce false or inaccurate information presented with apparent confidence. It is not an occasional bug, but a structural consequence of the current design of LLMs.

Harvest now, decrypt later: Strategy whereby state actors capture encrypted communications today with the intention of decrypting them when quantum computing matures. Present threat to data with long confidentiality lifetimes.

Hub & spokes: AI organizational model with a central Center of Excellence that establishes cross-cutting capabilities, while decentralized teams in lines of business develop specific solutions with cross functional reporting to the hub.

Humanoid robotics: Robots with human-like form and movement capabilities, integrated with AI models for perception, reasoning and learning.

Hyper-personalized phishing: AI-generated phishing attacks that analyze public profiles and corporate writing style to create highly personalized messages.

IRB (Internal Ratings-Based): Regulatory approach that allows banks to use internal models to calculate regulatory capital for credit risk.

Jagged intelligence: Andrej Karpathy's concept to describe the skill profile of today's LLMs: brilliant at complex tasks (mathematical olympiads) and fragile at seemingly simple tasks (comparing decimal numbers).

Jailbreak: Technique for bypassing the security controls of an AI model by means of instructions designed to make the system ignore its behavioral constraints.

LIME (Local Interpretable Model-agnostic Explanations): Explainability technique that generates local approximations of a complex model to explain individual predictions.

LLM (Large Language Model): Large-scale language model based on transformer architecture, trained on vast textual corpora. Technical foundation of part of today's generative AI. Examples: GPT, Claude, Gemini, Llama.

LLMOps (Large Language Model Operations): Specific MLOps extension to manage the properties of LLMs in production: non-deterministic behavior, prompts such as risk surface, hallucinations, cost per token and traceability of interactions.

LoRA (Low-Rank Adaptation): Efficient LLM fine-tuning technique that adapts the base model to a specific domain by modifying only a subset of parameters, drastically reducing the computational resources required.

Machine learning (ML): A branch of AI that develops algorithms capable of learning patterns from historical data without being explicitly programmed for each task. Unlike generative AI, classical ML models classify, predict and optimize; they do not generate content.

Machine unlearning: A set of experimental techniques that seek to selectively remove knowledge derived from specific data from an already trained model, in response to the GDPR right to be forgotten.

Many hands problem: Fuzzy liability problem in complex systems where damage occurs without deliberate intent and the causal chain is fragmented among multiple actors (designers, trainers, deployers, users).

MCP (Model Context Protocol): Open standard that standardizes how AI models interact with applications, data sources and external tools. It eliminates the technical debt of proprietary integrations and turns each tool into an asset reusable by any agent.

MLOps (Machine Learning Operations): Set of standardized processes and technology capabilities to reliably build, deploy and operationalize ML models throughout their lifecycle. It covers data preparation, experimentation, validation, deployment and monitoring.

Model drift: Silent degradation of the performance of a production model caused by changes in the distribution of input data relative to training data.

Multi-agent orchestration: Architecture in which multiple specialized AI agents collaborate under a central coordinator to achieve complex objectives, managing dependencies, priorities and information handoffs between agents.

Multimodality: Ability of an AI model to simultaneously process and generate multiple types of information (text, images, audio, video, code) in a single conversational architecture.

NIST AI RMF: National Institute of Standards and Technology's AI risk management framework. Organized into Govern-Map-Measure-Manage functions. Focused on trustworthy AI features (trustworthy AI).

Polymorphic malware: Malicious code that mutates its signature to evade detection. Generative AI powers it: advanced variants rewrite their code every 15 seconds while maintaining identical functionality, and defeat systems based on static signatures.

Prompt: A natural language instruction or input that a user provides to a generative AI system to elicit a response.

Prompt injection: Attack in which malicious instructions embedded in inputs (documents, URLs, external data) manipulate the behavior of the model to ignore constraints or execute unauthorized actions.

Quantum computing: Computational paradigm that exploits quantum properties (superposition, entanglement) to solve certain intractable problems for classical systems.

RAG (Retrieval-Augmented Generation): Architecture that complements an LLM with a real-time external information retrieval system. It reduces hallucinations by anchoring answers in verifiable documents, and separates updatable knowledge from the static memory of the model.

ReAct (Reason + Act): AI agent control loop that combines reasoning (situation analysis) and action (use of tools or APIs) iteratively. Foundation of the cognitive core of agentic systems.

Reskilling: Process of acquiring new competencies to perform professional roles different from the current ones, in response to the transformation of work by AI. Strategic lever given the structural imbalance between supply and demand for AI talent.

Shadow AI: Unauthorized use of generative AI tools in corporate environments, often on public platforms with no privacy guarantees.

SHAP (SHapley Additive exPlanations): Game theory-based explainability technique that quantifies the contribution of each variable to an individual prediction of a model.

SIEM / XDR / SOAR: Cybersecurity platforms: SIEM (Security Information and Event Management), XDR (Extended Detection and Response) and SOAR (Security Orchestration, Automation and Response).

Technoblocks: Partially incompatible spheres of technological influence (led by the US, China and Europe) with their own logics of governance, security and values.

Technological sovereignty: Ability of a state or organization to control the critical layers of the AI value chain (hardware, infrastructure, models, talent) in order to maintain strategic autonomy.

Transformer: Scalable neural network architecture based on attention mechanisms, introduced by Google in 2017. Technical foundation of all modern LLMs.

Turing test: Criterion proposed by Alan Turing (1950) to evaluate whether a machine exhibits intelligent behavior indistinguishable from that of a human in conversation.

UEBA (User and Entity Behavior Analytics): cybersecurity systems that establish dynamic baselines of normal behavior and detect anomalies.

Upskilling: Process of expanding existing competencies to adapt to new requirements of the same professional role, specifically in the context of AI adoption.

Vendor lock-in: Structural dependence on a single model or infrastructure vendor that makes migration difficult (rewriting integrations, compliance recertification, prohibitive costs). Strategic risk in AI adoption.

Vibe coding: Paradigm of software development by iterative natural language conversation with AI systems that interpret requirements, generate complete applications, detect errors and produce tests and documentation automatically. Term coined by Andrej Karpathy.

Our aim is to exceed our clients' expectations, and become their trusted partners

Management Solutions is an international consulting services company focused on consulting for business, risks, organization and processes, in both their functional components and in the implementation of their related technologies.

With its multi-disciplinary team (functional, mathematicians, technicians, etc.) of more than 4,000 professionals, Management Solutions operates through its 52 offices (22 in Europe, 24 in the Americas, 3 in Asia, 1 in Africa and 2 Oceania).

To cover its clients' needs, Management Solutions has structured its practices by sectors (Financial Institutions, Energy, Telecommunications and other industries) and by lines of activity, covering a broad range of skills – Strategy, Sales and Marketing Management, Risk Management and Control, Management and Financial Information, Transformation: Organization, Processes, & Technology, and New Technologies and Methodologies.

Javier Calvo
Partner and Chief AI Officer at Management Solutions
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán
Partner at Management Solutions
manuel.guzman@managementsolutions.com

Segismundo Jiménez
Director at Management Solutions
segismundo.jimenez@managementsolutions.com

Louis Perron
Manager at Management Solutions
louis.perron@managementsolutions.com

Management Solutions, Professional Consulting Services

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organization, technology and process-related advisory services.

For more information visit **www.managementsolutions.com**

Follow us at:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Wien Paris Bruxelles Amsterdam Copenhagen Oslo Stockholm Warszawa Wrocław Zürich Milano Roma Bologna Lisboa Beijing Abu Dhabi İstanbul Johannesburg Sydney Melbourne Toronto New York New Jersey Boston Pittsburgh Columbus Atlanta Birmingham Houston Phoenix Miami SJ de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Rio de Janeiro Lima Santiago de Chile Buenos Aires