AI Agents and Multi-Agent Systems in Highly Regulated Sectors

A Comparative Analysis of Europe, the United States, and China

2020–2026 | Second Edition | Revised & Expanded

reflexio

STRATEGIC DOSSIER

Issue No. 02 · Q1 2026 · v2.0

AI Agents and Multi-Agent Systems in Highly Regulated Sectors

A Comparative Analysis of Europe, the United States, and China

An evidence-based dossier for boards, regulators, risk officers and strategy leads in banking, energy and healthcare navigating multi-agent AI deployment under the EU AI Act, sector regulation in the United States, and China's coordination-first regime.

Publication date	Q1 2026
Publisher	RefleXio Intelligence
Series	RefleXio Intelligence, Strategic Dossier No. 02
Version	2.0, expanded comparative analysis, sector-by-sector evidence, regulatory architectures, failure-mode taxonomy, execution-readiness model
Prepared by	RefleXio Intelligence editorial desk

Independent intelligence on AI agents, regulation and execution in regulated sectors.

Table of Contents

0. EXECUTIVE SUMMARY

What this section means. AI adoption has reached critical mass globally, with 88% of organizations now use AI in at least one business function (McKinsey). But broad use is not the story. The story is the drop-off that follows. While roughly 79% of companies report active AI deployment, only 25% can point to measurable business results. That 54-point gap is not a technology problem. It is an organizational and governance problem, and it is worsening.

Why it matters. In highly regulated sectors such as banking, healthcare, and energy, the execution gap is not merely costly. It is a material risk event that spans regulatory enforcement, patient harm, infrastructure failure, and litigation exposure. Organizations that have deployed AI without the governance architecture to match are accumulating liabilities they have not yet quantified.

Implication. The next strategic investment in AI is not a better model. It is the governance infrastructure, process redesign, and accountability structures that convert AI capability into AI results. Organizations that make this investment now will hold durable competitive positions. Those that do not are building expensive pilots that will never reach production scale.

[Executive]

0.1 The Headline Is Real. It Is Also Misleading.

The period from 2020 to 2026 has produced a deceptively optimistic narrative about artificial intelligence. McKinsey's 2025 State of AI survey found 88% of organizations now using AI in at least one business function, up from 55% just two years prior. Generative AI moved faster than almost any prior technology wave. These numbers are real.

They are also misleading.

Beneath the headline figures lies a three-way divergence among the world's major AI powers, a divergence that reflects not merely different speeds of adoption but fundamentally incompatible models of what AI deployment actually means. And beneath the adoption figures in every region lies a convergent structural problem, namely the gap between what AI systems can do and what organizations are actually deploying in production at scale.

Operational Implication. Boards approving AI initiatives based on adoption metrics are approving the wrong indicator. The question is not "do we use AI?" Rather, it is "where does AI produce verifiable, measured returns?" Only 25% of organizations with active AI programs can answer that question affirmatively, per LinkedIn analysis of enterprise AI deployments.

[Executive]

0.2 Three Regions, Three Structural Conditions

Spain and the broader EU present the defining paradox of this analysis. Spain ranks sixth globally for workforce AI usage, with 41.8% of working-age Spaniards using AI tools at work in H2 2025, ahead of the United Kingdom and substantially ahead of the United States at 28.3%.

Yet at the enterprise level, the picture collapses. Eurostat data for 2024 places Spanish firm-level AI adoption at just 11.3%, below the EU-27 average of 13.5%.

In Spain's banking sector specifically, the evidence is more granular. The EBA, cited by Funcas, documents that approximately 60% of EU banks used AI in 2020; by 2024, that figure reached 86%, with 95% of banks running active AI pilots. A Funcas study of Spain's ten largest banking entities, a sample covering roughly 95% of sector assets, finds five of the ten scoring above 7/10 on an AI adoption index. Yet the Bank of Spain's Financial Stability Review notes that the majority of this AI use is concentrated in back-office and middle-office functions, namely risk management, compliance, and process automation, rather than autonomous client-facing decisions. High adoption. Low autonomy. That distinction is the execution gap in miniature.

The United States leads on every conventional investment metric. US private AI investment in 2024 reached $109.1 billion, nearly twelve times China's $9.3 billion. Financial services leads sectoral deployment, as 63% of US financial-sector workers used generative AI tools at work as of November 2025, per Federal Reserve data. The problem is depth, not breadth. Only 6% of organizations qualify as "GenAI high performers," meaning they attribute more than 10% of EBIT to AI deployment. Seventy-two percent of S&P 500 companies disclosed at least one material AI risk in 2025, up from 12% in 2023, per the Thomson Reuters Institute. The US is deploying fast and governing slowly. The gap between those two speeds is a structural liability.

China presents the most structurally distinct case. State-directed industrial policy has moved AI deployment from aspiration to mandate across prioritized sectors. Manufacturing AI penetration reached 25.9% in 2025 and China's "AI Plus" initiative targets 70% penetration across six key sectors by 2027. Yet the same state architecture that enables deployment at scale also constrains international transferability. Data localization requirements fragment training data; content moderation obligations create non-starters for EU and US deployment; and the IEA Energy and AI report documents how China's infrastructure-first approach to AI energy management creates a paradigm of massive AI-driven data center growth alongside ambitious grid optimization that remains structurally distinct from Western deployment models.

[Technical]

0.3 The Central Finding: Adoption Is Not Execution

A 2025 MIT Project NANDA study covering 300+ AI initiatives found that 95% of organizations deploying generative AI saw zero measurable P&L return. S&P Global Market Intelligence found that 42% of organizations abandoned most AI initiatives in 2025, up from 17% in 2024, a 147% increase in abandonment during the same period in which AI capability advanced most rapidly. Gartner predicts 40% or more of agentic AI projects will be canceled by 2027. Only 11% of agentic AI pilots reach production.

The failure modes are consistent across sectors, regions, and model generations. They include the absence of production-ready data infrastructure, the absence of defined business outcomes before build started, no named accountability for production quality, and no change management process for end-user adoption. None of these failures is a model problem. None is resolved by the next model release.

Healthcare's Evidence

In healthcare, SOTI's 2024/2025 Healthcare's Digital Dilemma report documents that 61% of global healthcare organizations used AI in 2024, rising to 81% in 2025. In Spain specifically, 37% used AI for personalized treatment, below the global average of 45%, and 49% for treatment planning. The binding constraint is explicit, as **58% of Spanish healthcare organizations cite legacy system limitations** as the primary obstacle to integrating AI, IoT, and telehealth. The AI is available. The organizational infrastructure to deploy it is not. The Spanish National Health System's eIASNS AI Strategy sets a roadmap for closing this gap, but the gap between national strategy and institutional deployment remains.

Energy's Evidence

In energy, the IEA's Energy and AI report frames a new paradigm in which AI systems are simultaneously the most powerful tool for energy grid optimization and a major new source of energy demand. AI data centers already account for over 10% of US electricity demand per IEA estimates, and that share is growing rapidly. The April 2025 Iberian blackout, affecting Spain and Portugal, demonstrated that growing automation complexity can actively impede operator response during cascade failures, per AI CERTs critical infrastructure analysis. The sector is deploying AI in smart grid management at pilot scale while simultaneously creating new grid stability risks through AI workload demand profiles that conventional management systems were not designed to handle.

> **Operational Implication. In all three sectors, the execution gap is not abstract. It takes sector-specific shapes. Banking** has high adoption locked in back-office functions; healthcare has organizational infrastructure that cannot support the AI tools being acquired; energy has advanced pilots stalled before industrial deployment while simultaneously creating new systemic risks from AI infrastructure demand.

> **KEY INSIGHT.** The question organizations need to stop asking is "Are we using AI?" Every competitor is. The question that determines outcomes is "Do we have the governance architecture, data infrastructure, process design, and accountability structures to deploy AI at production scale in a regulated environment?" For most organizations, the honest answer to that question determines the rest of the strategy.

[Executive]

0.4 Regulatory Divergence as Structural Architecture

Regulation is no longer an external constraint on AI deployment. It is now an architectural input that shapes what systems can be built, how they must be documented, what liability they carry, and who bears accountability when they fail.

The EU AI Act, entering full applicability in August 2026, creates the world's first comprehensive cross-sectoral AI legal framework with extraterritorial reach. For banking, the EBA's analysis of AI Act implications for the EU banking sector establishes that credit scoring, fraud detection, and customer-facing AI systems all fall within high-risk classifications requiring conformity assessment, audit trails, and EU database registration. This is not a compliance overhead. It is a design specification. Organizations building AI systems for EU banking without understanding the EBA's Act mapping are building systems they will need to fundamentally reconstruct.

The BIS Financial Stability Insights report on AI regulation adds a systemic dimension, as agentic AI in financial markets creates emergent risks, including herding behavior, involuntary collusion between models operating on similar logic, and coordinated cyberattack surfaces, that individual institution governance cannot resolve. These are sector-level systemic risks requiring regulatory coordination that is still forming. Organizations deploying multi-agent systems in financial markets are operating ahead of the governance frameworks designed to contain their systemic risk profile.

> **Operational Implication.** Any organization building or procuring AI systems for EU-regulated markets must treat regulatory compliance as an architectural requirement, not a pre-launch review. The compliance cost per model runs approximately €29,277 annually; the cost of reconstructing non-compliant architecture post-deployment is orders of magnitude higher.

[Executive]

0.5 What Is at Stake: The Three Decisions

This document is organized around three categories of decisions facing boards and C-suites in 2026.

Decision 1, Investment allocation. The dominant AI investment error is continued spending on model acquisition and pilot initiation in organizations that have not first built the governance, data, and process infrastructure to extract value from what they already have. Wharton research cited in Trax Group's analysis found organizations underestimate production deployment complexity by 300–500%. The highest-return AI investment for most boards in 2026 is not a new model. It is closing the governance and data readiness deficit that is preventing existing investments from generating returns.

Decision 2, Risk quantification. The organizations deploying AI in regulated sectors without commensurate governance are accumulating liabilities they have not quantified, including regulatory enforcement exposure (under the AI Act, GDPR, MiFID II, MDR), litigation risk (algorithmic bias, patient harm, market manipulation), and operational risk (cascade failures, data breaches, model drift). The 89% of enterprises lacking formal AI governance frameworks are not just operationally inefficient. They are carrying risk that boards have implicitly accepted without explicitly pricing.

Decision 3, Competitive positioning. The organizations that build production-quality AI governance infrastructure now, including audit trails, explainability outputs, human-in-the-loop protocols, and accountability structures, are building a "compliant by design" architecture that becomes a competitive differentiator as AI Act enforcement tightens. The DECESOSIA multi-agent insurance system operating in Spain's financial sandbox and CaixaBank's agentic AI deployment for customer product contracting represent the frontier of what "compliant by design" agentic AI looks like in practice. These are not experiments. They are templates for the next generation of regulated AI deployment.

The chapters that follow document the evidence base for each of these decisions, namely the execution gap in structural terms, the sector-by-sector manifestations, the regional divergence in approach, and the organizational conditions under which AI deployment succeeds. The analysis is precise about what the data shows and direct about what it implies.

1. KEY STRATEGIC SIGNALS

What this section means. Six structural conditions define where AI deployment succeeds and fails in regulated sectors. These are not predictions. They are patterns that have already played out and are determining competitive outcomes now.

Why it matters. Each signal is a decision point. Organizations that have correctly read the signal are building durable positions. Those that have misread it are accumulating costs, risks, and lost ground that compound quarterly.

Implication. This chapter is written for executives who will not read the rest of the document. The six signals below, taken together, constitute the strategic brief. Every chapter that follows is evidence for one or more of these claims.

[Executive]

The following six signals are not summaries of findings. They are claims, each with direct bearing on strategic decisions that boards and C-suites will make in 2026 and 2027. They are written to provoke disagreement, because the organizations that reject them today are the ones that will be explaining their results three years from now.

1. The execution gap is a structural problem, not a temporary lag. Adoption metrics are now a liability indicator, not a progress indicator.

Seventy-nine percent of companies report active AI adoption. Only 25% report measurable business results, per LinkedIn's enterprise AI deployment analysis. The gap between these two figures, 54 percentage points, has not closed as models improved. It widened. The S&P Global abandonment rate jumped from 17% in 2024 to 42% in 2025, a 147% increase during the same period in which AI capability advanced most rapidly. This is not a temporary lag awaiting better technology. It is a structural condition produced by organizations that have invested in model acquisition without building the governance, data, and process infrastructure required to use models in production. Every board that approves an AI initiative without first demanding evidence of data readiness, governance architecture, and named accountability for results is approving a pilot that will not graduate.

2. High adoption, low execution is the default condition across all regulated sectors. The sector-specific shapes of the gap define different intervention strategies.

Banking in Europe leads on raw adoption, with 86% of EU banks using AI and 95% running active pilots. But the Bank of Spain and EBA confirm that use is concentrated in back-office and middle-office functions, not in autonomous decision-making or customer-facing agentic deployment. Healthcare globally moved from 61% to 81% AI adoption in a single year, yet 58% of Spanish healthcare organizations cite legacy infrastructure as the primary obstacle to deployment. Energy has advanced technical pilots in smart grids while the IEA documents that AI's own infrastructure demand is creating a new category of grid management risk. In each case, the intervention required differs. Banking needs governance frameworks to unlock autonomous deployment; healthcare needs infrastructure investment to enable integration; energy needs coordination protocols between AI operators and grid managers.

3. Regulation is now a design force, not a compliance overhead. Organizations that treat it as the latter will require expensive reconstruction, while those that treat it as the former will hold structural competitive advantage.

The EU AI Act is not an obstacle to AI deployment. It is a specification for how AI systems must be built to operate sustainably in European markets. The EBA's AI Act banking sector analysis maps specific obligations for credit scoring, fraud detection, and customer-facing systems, obligations that are architectural, not cosmetic. Spain's financial sandbox is already testing what compliant agentic deployment looks like in practice, with projects like DECESOSIA designing multi-agent insurance systems with full regulatory accountability from the ground up. The organizations building "compliant by design" architectures today are building systems that will not require reconstruction when the Act reaches full applicability in August 2026. Those treating compliance as a post-build review are building technical debt that will cost more to resolve than it would have cost to prevent.

4. Multi-agent systems are a risk multiplier, not just a capability multiplier. The governance overhead scales combinatorially while most organizations' governance frameworks remain unchanged from single-model deployments.

Galileo AI's analysis of production multi-agent failures documents the mathematics, showing that 4 agents create 6 potential failure interaction points, 10 agents create 45, and 20 agents create 190. The BIS FSI Insights report elevates this from a firm-level to a systemic concern, as multi-agent deployment across financial institutions sharing similar model architectures creates herding behavior, involuntary collusion risk, and coordinated cyberattack exposure that individual institution governance cannot contain. Deloitte's agentic AI risk analysis for banking now recommends agent-specific governance frameworks with defined role identities, unified logging, kill-switch protocols, and adversarial testing, capabilities that most organizations have not built for their existing single-model deployments, let alone for multi-agent systems.

5. Geopolitical divergence in AI governance is structural, not transitional. Organizations operating across EU, US, and Chinese markets must maintain architecturally distinct AI pipelines, and that cost is permanent.

The EU mandates explainability, audit trails, and conformity assessment for high-risk AI. The US favors speed over precaution, delegating governance to voluntary frameworks and sector regulators. China requires state-aligned content moderation and data localization that prevents cross-border deployment of Chinese AI systems. These are not converging. An organization deploying AI in credit scoring across all three markets operates three different compliance regimes simultaneously. The €29,277 annual compliance cost per model in the EU is a floor, not a ceiling. Multi-model deployments multiply it. The organizations that have not yet modeled the full regulatory cost of their AI portfolio across geographies are underestimating total cost of ownership by a significant margin.

6. Spain holds a first-mover position in regulated agentic AI that is currently underutilized. The combination of EU regulatory leadership, advanced banking sector AI adoption, and active sandbox experimentation creates a template export opportunity that requires deliberate organizational investment to capture.

CaixaBank's generative AI agent, which guides customers through full product contracting within the mobile app, represents the frontier of customer-facing agentic banking in Europe. The DECESOSIA project in Spain's financial sandbox is one of the first formally regulated multi-agent systems for insurance management in Europe, covering policy contracting, portfolio optimization, claims management, and omnichannel communication under direct regulatory oversight. These are not isolated experiments. They are proof-of-concept deployments for a "compliant by design" agentic AI model that has no equivalent at production scale in the US or China. The strategic question for Spain is whether these early positions are institutionalized into sector-wide capability or remain isolated examples while the organizational infrastructure to replicate them goes unbuilt.

2. THE EXECUTION GAP

What this section means. The execution gap is the distance between what AI systems can demonstrably do and what organizations are actually deploying in production at scale. This chapter defines that gap with precision, identifies its four structural root causes, documents the statistical evidence for its magnitude, and shows, sector by sector, how it manifests differently in banking, healthcare, and energy.

Why it matters. Understanding the gap is the precondition for closing it. Organizations that treat AI underperformance as a model problem will keep purchasing new models and producing the same results. Organizations that treat it as an organizational readiness problem can close it with the right interventions.

Implication. The gap is not primarily a technology challenge, and it is not resolved by model upgrades. It is closed by investment in governance architecture, data infrastructure, process redesign, and cultural readiness, in that order of organizational priority.

[Executive]

2.1 Defining the Gap: Capability vs. Deployment

The execution gap did not exist when AI was an experimental discipline. It became a strategic problem at the moment AI capability crossed the threshold of genuine usefulness, approximately 2022–2023 for generative AI, and earlier for machine learning in specific domains like fraud detection and predictive maintenance. That crossing meant the bottleneck shifted from "can the technology do this?" to "can the organization deploy, govern, and sustain this?"

Research from Anthropic analyzed by Ability.ai shows that modern foundation models can handle the vast majority of tasks in coding, finance, legal, and administrative work at a level competitive with skilled human professionals. The gap between this capability and actual enterprise production deployment is not a function of model inadequacy. It is a function of integration complexity, data readiness, governance absence, and organizational structures designed for human decision-making that have not been redesigned for human-AI collaboration.

[Technical]

The Pre-2022 and Post-2022 Conditions

Before 2022, the AI deployment question was primarily technical: could models perform reliably enough for production use? That question has been answered affirmatively across a broad range of applications. The post-2022 question is organizational: can institutions build the surrounding infrastructure to deploy, govern, monitor, and sustain AI systems in production? That question is being answered negatively in the overwhelming majority of cases.

This shift matters for diagnosis. Organizations still running pre-2022 AI strategy, focused on model capability, benchmarks, and "build vs. buy" model decisions, are solving the wrong problem. The solved problem is whether capable models exist. The unsolved problem is whether the organization can use them.

> **Operational Implication.** Any AI initiative that begins with model selection rather than with governance architecture, data audit, and process redesign mapping is organized for pre-2022 conditions. The scoping methodology, not the model choice, determines deployment outcomes.

[Technical]

2.2 The Four Structural Deficits

The execution gap can be operationalized as four distinct organizational readiness deficits. Each must be present for deployment to succeed. They are not substitutable: strength in one dimension does not compensate for weakness in another.

Deficit 1: Governance

[Executive]

Does the organization have defined AI ownership, accountability structures, audit mechanisms, and escalation protocols? Can it answer "who is responsible when this system makes an error?"

Epicenter's 2025 AI Governance Benchmark Report found that **89% of enterprises lack a formal AI governance framework**. Without governance readiness, no AI system, regardless of capability, can operate in a regulated environment. Governance failures produce specific downstream effects. Systems drift from their training distribution without detection; models produce systematically biased outputs that no one is positioned to identify; and regulatory inquiries arrive without the audit trail required to demonstrate compliance.

Deloitte's State of AI in the Enterprise 2026 found 58% of leaders identify disconnected governance systems as the primary obstacle preventing AI scaling. This is not a philosophy problem. It is an organizational design problem.

> **Operational Implication.** Before any AI initiative receives production approval, the organization must answer several questions. Who owns this system's behavior? What triggers an escalation? What is the audit mechanism? What is the kill-switch process? If these questions lack named, specific answers, the system is not production-ready regardless of its technical performance.

[Technical]

Deficit 2: Coordination

In multi-agent and cross-functional AI deployments, coordination failure is the second structural deficit. AI model deployment operates at software speed; governance review operates at committee speed. Data science teams in forward-leaning enterprises ship new models weekly; governance processes, when they exist, run on monthly or quarterly cycles. This timing mismatch means governance reviews consistently arrive after deployment, creating retroactive compliance rather than proactive oversight.

Accountability diffusion compounds the problem. In most organizations, no single function owns AI model behavior end-to-end. Legal focuses on regulatory exposure; IT focuses on infrastructure security; operations focuses on output metrics; risk focuses on financial loss quantification. As a result, nobody owns model behavior, and no one is positioned to identify when an AI system is drifting from acceptable performance. BIS analysis extends this to the systemic level, showing that in banking, multi-agent systems operating across institutions with similar architectures create coordination risks, including herding, involuntary collusion, and coordinated vulnerability exposure, that exceed any individual institution's governance perimeter.

> **Operational Implication.** For any AI system operating across functional silos or involving multiple agents, a single named accountability owner must exist, not a committee. Committees coordinate; they do not own. The owner must have authority to pause or terminate the system and accountability for its production behavior.

[Technical]

Deficit 3: Architecture

[Executive]

Does the technical architecture of the AI system support audit, explainability, and integration with existing business processes? Legacy processes built around sequential human decision-making, such as loan approval workflows, clinical referral pathways, and grid dispatch protocols, are structurally incompatible with AI agent operation.

The most common failure mode documented in the Digital Applied March 2026 survey of 650 enterprise technology leaders is integration complexity with legacy systems, a process and architecture readiness failure, not a model failure. Organizations scope AI initiatives as model selection decisions, then discover that process redesign and architecture integration were the actual project.

Informatica's CDO Insights 2025 found 43% of organizations cite data quality and readiness as the primary obstacle to AI success, the single largest factor across all failure categories. Production AI requires data that is current, accurately labeled, consistently formatted, and integrated across the organizational silos that AI workflows need to bridge. The ability to access data is not the same as the ability to use it for AI deployment.

> **Operational Implication.** The architecture investment for AI deployment must include data pipeline production-readiness, audit trail generation in every decision step, explainability outputs for regulated decisions, and integration design that accounts for legacy system constraints. These are not post-build additions. They are foundational design requirements.

[Technical]

Deficit 4: Organization

The fourth structural deficit is organizational culture, that is, whether end users and domain experts will engage constructively with AI outputs, or route around the system, selectively apply results, or abandon the tool when it makes errors.

In a 2024 Beth Israel Deaconess study covered by The New York Times, doctors showed a consistent tendency to stick with their own diagnoses even when AI offered more accurate alternatives. Camunda survey data found 80% of organizations cited lack of transparency as a barrier to agentic AI progress, a cultural readiness problem, not an explainability technology problem.

Operational Implication. End-user co-design is not optional. AI tools built without systematic involvement of domain experts in workflow design will be adopted superficially and abandoned within months. The investment in change management and user co-design should equal or exceed the investment in model selection.

KEY INSIGHT. Organizations that have closed the execution gap did not do so by finding better AI models. They did it by building production-quality data infrastructure before deployment, defining measurable business outcomes before build started, co-designing workflows with end users, and naming a single individual accountable for production quality. None of these success factors is a technology decision.

[Executive]

2.3 Evidence of the Gap

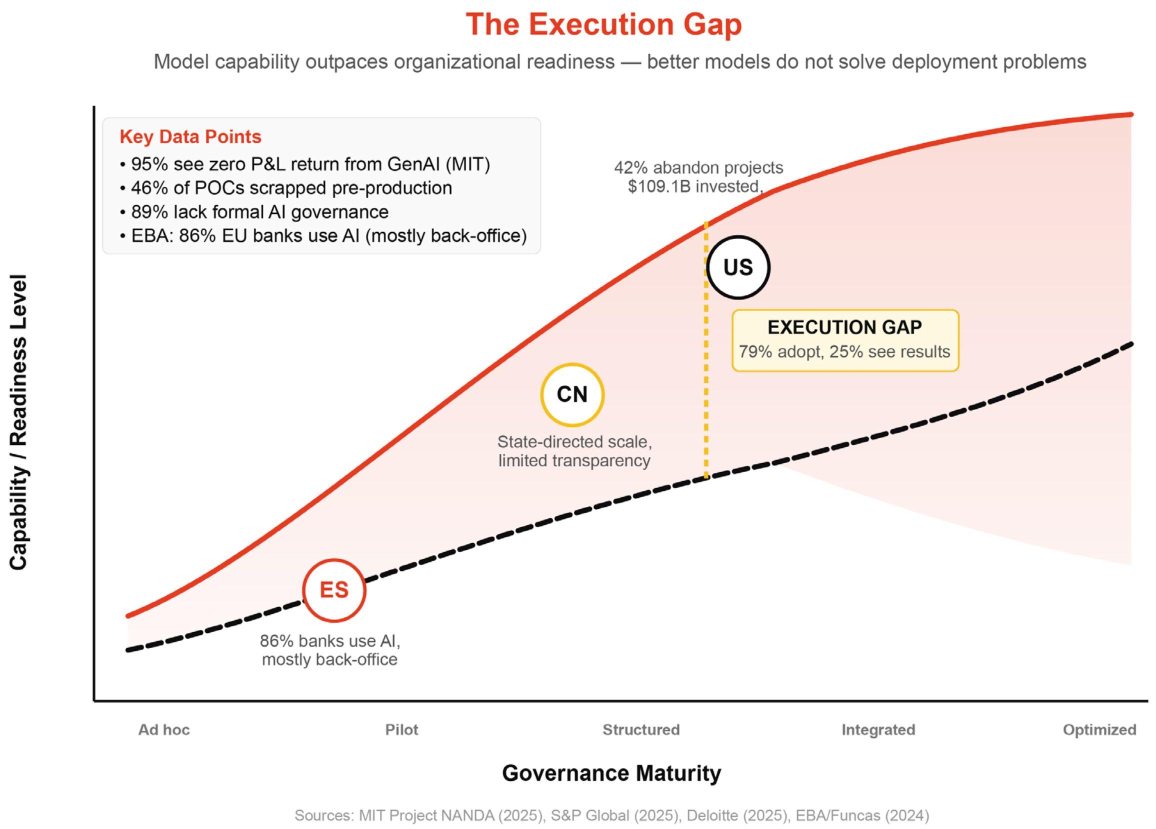

Figure 1. The AI Execution Gap — Capability vs. Production Deployment

The statistical evidence for the execution gap is precise, convergent, and sector-agnostic.

Adoption vs. Results

Metric	Statistic	Source
Organizations reporting active GenAI adoption	79%	LinkedIn Enterprise AI Analysis
Organizations achieving measurable business results	25%	LinkedIn Enterprise AI Analysis
Execution gap (adoption minus results)	54 percentage points	—
AI initiatives abandoned in 2025	42%	S&P Global / CIO Dive
AI initiatives abandoned in 2024	17%	S&P Global / CIO Dive
Agentic AI pilots projected to be canceled by 2027	40%+	Gartner, via Squirro
Agentic AI pilots crossing into production	11%	Process Excellence Network
Organizations with zero measurable P&L return from GenAI	95%	MIT Project NANDA
Enterprises lacking formal AI governance framework	89%	Epicenter 2025 Benchmark

Strategic Implication. The gap between adoption and results is not narrowing. It is widening as deployment accelerates ahead of governance. Organizations cannot close the gap by adopting more AI; they close it by building the organizational infrastructure that converts AI capability into AI results. The data is unambiguous that this infrastructure is the scarce resource, not the AI models themselves.

[Technical]

Why Better Models Do Not Solve the Problem

MIT Project NANDA examined 300+ AI initiatives and found the failure modes include the absence of defined business outcomes before build, data infrastructure not built for production, misaligned incentives with no end-user co-design, the absence of defined success metrics, and a governance and ownership vacuum. None of these failures is resolved by a model upgrade.

The McKinsey 2024 State of AI Survey identified the "GenAI paradox," in which rapid technological breakthroughs coexist with slow productivity gains, directly analogous to prior technology disruptions where adoption preceded the organizational redesign required to generate returns. Of 876 surveyed companies, only 46, approximately 5%, were identified as GenAI high performers. They did not succeed because they had better models. They succeeded because they had better execution frameworks.

A further dimension is that organizations overestimate the stability of model performance in production conditions. Distribution shift, the divergence between training data statistics and live production input statistics, is documented as a leading cause of production AI failure. A model achieving 92% accuracy on benchmark datasets will consistently underperform in production because the tail of the input distribution, that is, rare, malformed, ambiguous, or adversarial inputs comprising 1 to 5% of production volume, is never adequately tested in controlled pilots. In regulated sectors, this tail is precisely where AI causes its most consequential errors.

> **Operational Implication.** Performance benchmarks should not be presented to boards as evidence of production readiness. A system must pass production simulation tests, using live data distributions, stress inputs, and adversarial edge cases, before it qualifies for regulated-sector deployment approval.

[Executive]

2.4 From Adoption to Execution: Sector Evidence

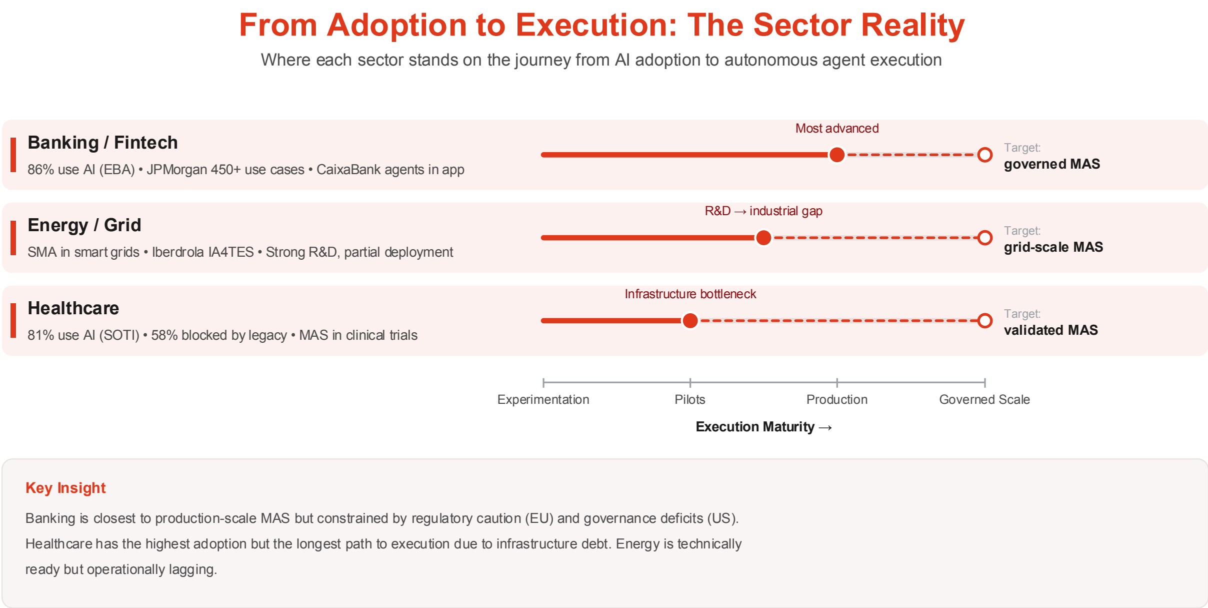

Figure 2. From Adoption to Execution — The Transition Framework

The execution gap is not sector-agnostic. In highly regulated environments, the gap shifts from operational underperformance to material risk events, including regulatory enforcement, patient harm, infrastructure failure, and litigation exposure. Three sectors document this most sharply, and each illustrates a different shape of the same fundamental problem.

[Technical]

Banking: High Adoption, Low Autonomy

The Adoption Picture

The EBA, cited by Funcas, documents the trajectory clearly. EU banks using AI grew from approximately 60% in 2020 to 86% in 2024, with 95% of banks running active pilots when ongoing programs are included. This is not marginal adoption. It is a near-universal presence.

In Spain specifically, Funcas research on Spain's ten largest banking institutions found that all ten score above 5/10 on an AI adoption index, with five exceeding 7/10, performance consistent with the highest-adopting cohort across Europe.

Where the Execution Gap Appears

Despite this adoption breadth, the Bank of Spain's Financial Stability Review is explicit about where the deployment is concentrated, namely risk management, compliance automation, fraud detection, regulatory reporting, and back-office process optimization. GenAI and RAG architectures are deployed predominantly in analyst support roles, not in autonomous customer-facing decisions or fully autonomous operational workflows. The reason is not technical inadequacy. It is regulatory caution, as AI Act risk classifications for credit scoring and customer-facing AI create compliance requirements that most institutions have not yet built the governance infrastructure to satisfy.

The practical consequence is a population of EU banks that have adopted AI extensively but deployed it narrowly, using capable technology in the safest possible functions rather than the highest-value ones. This is rational risk management in the short term. It is the execution gap in the medium term, as the organizational capability to operate AI in production at the functions where it creates most value, such as credit decisions, customer advisory, and portfolio management, remains underdeveloped because those functions have not been the deployment priority.

Agentic AI at the Frontier: Spanish Examples

Two Spanish examples illustrate what moving beyond this constraint looks like in practice.

CaixaBank has deployed generative AI agents in its mobile banking application, agents that explain products, compare alternatives, and accompany customers through the full product contracting process including consumer credit. The architecture is agentic, as a single agent manages multi-turn dialogue, queries internal product and compliance systems, and orchestrates contracting steps autonomously. This is not formally described as a multi-agent system, but it is structurally agentic, and it represents the first customer-facing agentic banking deployment at scale in Spain, operating in a regulated environment with EU consumer protection obligations.

The DECESOSIA project, operating within Spain's financial regulatory sandbox under oversight of the Dirección General de Seguros, proposes a genuinely multi-agent system for life insurance policy management, with autonomous agents handling policy contracting, portfolio optimization, claims processing, and omnichannel customer communication, operating under full regulatory accountability. DECESOSIA is one of the first formally regulated multi-agent insurance systems in Europe, a pioneering case of the governance-first architecture that the EU AI Act will eventually require from all high-risk deployments.

> **Operational Implication.** Spanish banking and insurance institutions have proof-of-concept evidence, at regulatory production scale, that agentic AI can operate within EU compliance frameworks. The execution challenge is not demonstrating that it can be done; it is building the organizational infrastructure to replicate these deployments at institutional scale. That requires governance frameworks, agent identity and logging architecture, human-in-the-loop protocols, and regulatory engagement processes, the investments that separate 86% adoption from production-scale agentic deployment.

[Technical]

Energy: Advanced Pilots, Partial Deployment

The IEA Paradigm Shift

The IEA's Energy and AI 2025 report frames the energy sector's relationship with AI in terms of a structural paradox, as AI is simultaneously the most powerful available tool for energy grid optimization, renewable integration, and demand-side management, and the fastest-growing source of new energy demand. AI data centers and compute infrastructure already represent over 10% of US national electricity demand per IEA estimates; this figure is growing rapidly as generative AI and agentic workloads scale.

This creates a sector-specific version of the execution gap, as energy organizations are deploying AI to manage a grid whose stability is increasingly challenged by AI infrastructure demand profiles. The April 2025 Iberian blackout, affecting Spain and Portugal, illustrated the consequence. Growing automation complexity in grid management systems actively impeded operator response time during the cascade failure. Human operators, increasingly dependent on automated systems for routine grid management, had diminished capacity to intervene manually when automation failed. This is not a hypothetical risk. It is a documented failure mode.

The Research-to-Production Gap

European academic and applied research on multi-agent systems for energy is substantial. Comprehensive reviews published in Applied Energy and CSEE Journal document multi-agent systems applied to smart grid management, distributed energy resource coordination, microgrid optimization, and demand response. The European Parliament's briefing on AI and the energy sector acknowledges AI's role in grid management explicitly and anticipates a Strategic Roadmap for AI in Energy Digitalization in 2026.

In Spain, multi-agent research and pilot projects are active: a Spanish university consortium produced an AI-Driven Multiagent IoT System for Energy Consumption Anomaly Detection, and IA4TES documents Spanish research programs on multi-agent intelligent energy market management. These are technically sophisticated. They are not yet deployed at industrial production scale.

The structural gap in energy is the distance from research-grade pilot to operational industrial deployment. The barriers are specific. Grid operators need coordination protocols with AI system operators before autonomous AI load management can be deployed safely; cybersecurity requirements for multi-agent systems in critical infrastructure are more demanding than for conventional SCADA systems; and the AI Act's critical infrastructure provisions apply to energy management AI, creating compliance requirements that most pilot programs have not yet navigated.

> **Operational Implication.** Energy sector organizations with active AI pilots should be designing the governance and coordination infrastructure for industrial deployment now, not after the pilot concludes. The coordination protocols with grid operators, the cybersecurity testing regime for multi-agent systems, and the AI Act compliance pathway must be developed in parallel with the technical pilot, or the gap between pilot completion and production deployment will extend by 18–24 months minimum.

[Executive]

Healthcare: Growing Usage, Blocked by Infrastructure

The Adoption Trajectory

SOTI's Healthcare's Digital Dilemma report documents healthcare AI adoption rising from 61% of organizations globally in 2024 to 81% in 2025. Within a single year, healthcare moved from majority adoption to near-universal adoption. In Spain, the data is more nuanced. 37% of healthcare organizations use AI for personalized treatment (versus 45% globally) and 49% for treatment planning (slightly above the 46% global average).

These figures suggest Spain is close to the global average on AI usage breadth, but the gap on personalization specifically is revealing. Personalized treatment is the highest-complexity AI healthcare application, requiring integration across patient history systems, genomics data, treatment outcome databases, and clinical evidence repositories. The lower penetration rate in Spain for this specific application reflects the infrastructure constraint precisely.

The Legacy System Barrier

SOTI's research on Spain is explicit. **58% of Spanish healthcare organizations identify legacy IT system limitations as the primary obstacle to integrating AI, IoT, and telehealth**. This is not a problem of AI tool availability, since the tools exist and are being acquired. It is a problem of integration infrastructure, namely the clinical data repositories, patient identity systems, encounter records, lab value databases, medication reconciliation platforms, and imaging archives that AI clinical decision support requires, which are maintained in separate, incompatible legacy systems with different update schedules, different data standards, and different governance structures.

Spain's National Health System eIASNS AI Strategy directly addresses this. The strategy establishes a roadmap for AI deployment in diagnostic imaging, risk stratification, prescription support, and clinical logistics, always under principles of security, equity, transparency, and human supervision. The strategy is architecturally sound. The gap is implementation, as the 58% of organizations blocked by legacy systems will require substantial infrastructure investment before the eIASNS strategy translates into production AI deployment.

The Multi-Agent Frontier in Healthcare

At the global research frontier, a scoping review of AI agents in healthcare research identifies three established archetypes, namely clinical decision support agents, care pathway and operations management agents, and biomedical research agents. Research comparing single-agent and multi-agent healthcare AI finds that orchestrated multi-agent systems outperform single agents in complex tasks, specifically where different agents specialize in history summarization, evidence retrieval, risk assessment, and treatment plan generation, with an orchestrator integrating outputs.

In Spain and Europe, these systems are predominantly at pilot or research stage, constrained by the intersection of regulatory density, given that a European Parliament mapping of EU and 10-country healthcare AI regulation identified at least 141 binding norms across medical device regulation, data protection, and liability frameworks, and the legacy infrastructure barriers documented above.

> **Operational Implication:** For Spanish healthcare organizations, the highest-return near-term AI investment is infrastructure, not application. Until the legacy integration barriers are addressed, AI tool acquisition will continue to produce the pattern documented in the SOTI data:

growing adoption of AI tools for lower-complexity applications (administrative automation, treatment planning support) alongside persistent inability to deploy the high-value applications (personalized treatment, multi-source clinical synthesis) that require integrated data infrastructure. The eIASNS strategy provides the architectural framework; institutional boards must fund the infrastructure investment required to execute it.

KEY INSIGHT. In banking, the execution gap looks like 86% adoption concentrated in back-office functions rather than the autonomous customer-facing deployments that generate competitive advantage. In energy, it looks like technically sophisticated pilots that have not crossed into industrial production. In healthcare, it looks like AI tool adoption blocked by the legacy systems that AI needs to connect to be useful. Same gap, three different shapes, requiring three different interventions.

[Deep Dive]

2.5 The Systemic Risk Dimension of the Execution Gap

The preceding analysis focuses on organizational execution failures. There is a further dimension that boards must understand, namely that the execution gap in multi-agent and agentic AI systems creates systemic risks that exceed individual organizational liability.

The FSB's Financial Stability Implications of Artificial Intelligence identifies the mechanism precisely. As financial institutions deploy AI systems with similar underlying architectures, such as similar training data sources, similar model families, and similar optimization objectives, the behavior of these systems under stress conditions becomes correlated. During market stress, correlated AI responses can amplify price movements rather than moderate them (herding), and multiple systems with similar vulnerability profiles create a coordinated attack surface for adversarial exploitation.

The BIS FSI Insights report extends this analysis. In agentic AI deployments where multiple institutions use similar model architectures, the risk of involuntary collusion, where agents independently optimize for similar objectives and produce market outcomes equivalent to coordinated action, cannot be addressed by individual institution governance alone. This is a regulatory gap that is structural, not transitional.

For energy, the parallel risk is grid stability. As more grid management functions are automated using AI systems with correlated response logic, the risk of simultaneous failure during abnormal grid conditions increases. The Virginia data center cascade of 2024, where sixty AI data center facilities triggered protective disconnection logic simultaneously within 82 seconds, documents the production manifestation of this risk.

Operational Implication. For regulated financial institutions, the systemic risk dimension of multi-agent AI requires engagement with regulators, including the EBA, Banco de España, and the ECB, on governance frameworks that address herding and collusion risk at the sector level, not just at the institution level. Waiting for regulatory guidance to arrive is not a sufficient risk management posture; proactive engagement on sector-level governance frameworks is a fiduciary responsibility.

References — Chapters 0, 1, and 2

Banking and Financial Services

- Banco de España — «La inteligencia artificial en el sistema financiero» (Financial Stability Review 47, 2024)
- Funcas — «La inteligencia artificial en la banca europea: adopción y casos de uso» (Nota OFT-2/2025)
- Funcas / Carbó et al. — «Adopción de la inteligencia artificial en la banca» (Papeles de Economía Española 186)
- Funcas — «La adopción de la nueva tecnología bancaria en España y Europa»
- CaixaBank develops an AI agent to help customers explore products in the app
- CaixaBank launches the first AI agent that assists customers in purchasing products from its app
- Ministerio de Asuntos Económicos y Transformación Digital — «Informe Bienal del Sandbox Financiero 2024–2025» (Spanish Financial Sandbox, Biennial Report)
- AI Act: implications for the EU banking and payments sector — EBA
- Regulating AI in the financial sector: recent developments and main challenges — BIS FSI Insights
- The Financial Stability Implications of Artificial Intelligence — FSB
- Managing the new wave of risks from AI agents in banking — Deloitte
- Artificial Intelligence in Financial Services — World Economic Forum 2025
- GARP — AI and Governance Gaps in Financial Services (2024 Benchmarking Survey)
- Energy
- Energy and AI — IEA Report 2025
- AI and the energy sector — European Parliament Briefing (2025)
- Multi-agent systems applied for energy systems integration: State-of-the-art — Applied Energy
- Comprehensive Overview of Multi-agent Systems for Controlling Smart Grids — CSEE Journal
- AI-Driven Multiagent IoT System for Energy Consumption Anomaly Detection — RUA (University of Alicante)
- Mercados energéticos Inteligentes — IA4TES
- World Energy Outlook 2024 — IEA (Executive Summary)
- Critical Infrastructure Failure: AI Grid Risks and Response (Iberian Blackout, Virginia Cascade) — AI CERTs
- AI Data Center Electricity Demand and Grid Stability — arXiv (September 2025)

Healthcare

- Healthcare's Digital Dilemma — SOTI Industry Report 2024/2025
- Ministerio de Sanidad — «Estrategia de Inteligencia Artificial para el SNS (eIASNS)» (Spanish National Health System AI Strategy)
- Mapping the regulatory landscape for artificial intelligence in health in the EU — npj Digital Medicine
- Artificial intelligence agents in healthcare research: A scoping review
- Orchestrated multi-agent AI systems outperform single agents in health care — Medical Xpress
- Hospital Predictive AI Adoption 2023–2024 — ONC/HealthIT.gov
- TruDi AI Surgical Navigation Investigation — Reuters (February 2026)
- IA Revoluciona la Sanidad en España — Takeda

Execution Gap and Governance

- McKinsey State of AI 2025 — "AI at Work but Not at Scale"
- MIT Project NANDA — Why 95% of Corporate AI Projects Fail
- S&P Global Market Intelligence / CIO Dive — AI Project Abandonment Rates
- Gartner — 40% of Agentic AI Projects to be Canceled by 2027
- Process Excellence Network — Why Most Agentic AI Pilots Fail (11% crossing to production)
- LinkedIn — AI Execution Gap: 79% Adoption, 25% Results
- Epicenter — AI Governance Framework Gap (2025 Benchmark Report)
- Digital Applied — AI Agent Scaling Gap Survey (March 2026, 650 enterprise technology leaders)
- Galileo AI — Why Multi-Agent Systems Fail
- McKinsey State of AI 2024 (GenAI Paradox, 5% high performers)
- Wharton / Trax Group — Gartner 2025 Hype Cycle GenAI Analysis (300–500% underestimation)
- AI Adoption Gap and Workflow Deployment Bottleneck — Ability.ai
- WorkOS Blog — Why Most Enterprise AI Projects Fail (Informatica CDO Insights 2025)
- Why Healthcare AI Fails — TATEEDA

Regulatory and Geopolitical

- Eurostat Enterprise AI Adoption 2024 — 13.5% EU Average
- Bank of Spain EBAE Survey — AI Adoption in Spanish Firms
- Stanford HAI AI Index 2025 — Economy Chapter (US $109.1B investment)
- Federal Reserve FEDS Notes — Monitoring AI Adoption in US Economy (63% financial sector)
- Thomson Reuters Institute — AI Governance Gap and ESG Risks
- EU AI Act Compliance Cost Statistics — SQ Magazine
- Spain 41.8% AI Workforce Adoption — Microsoft Global AI Adoption Report via Idealista
- New York Times — ChatGPT and Doctors' Diagnoses (Beth Israel Deaconess Study)
- Banco de España and BSC Sign Agreement to Promote AI in Financial Sector
- Cognizant — GenAI Adoption in Spain (22% below global momentum)
- China "AI Plus" Initiative — USSC

- End of Chapters 0–2. Chapters 3 onwards cover country-by-country analysis (Europe, United States, China), sector deep-dives (Banking, Healthcare, Energy), and implementation frameworks.

3. Global Landscape of AI Agents (2020–2026)

What this section means. Three distinct regions, Europe, the United States, and China, are deploying AI in fundamentally different ways, at different speeds, and toward different ends. The headline adoption numbers are largely comparable; the underlying architectures are not.

Why it matters. Executives who treat AI adoption as a single global trend systematically misread competitive positioning. A 60% banking adoption figure in the EU and an 83% industrial deployment figure in China measure different phenomena, fund different capabilities, and produce different risks.

Implication. Before benchmarking your organization's AI maturity against regional or sector averages, identify which adoption metric you are using, and whether it captures the agentic, multi-step deployments that will define competitive differentiation through 2030.

3.1 Adoption vs. Deployment Reality

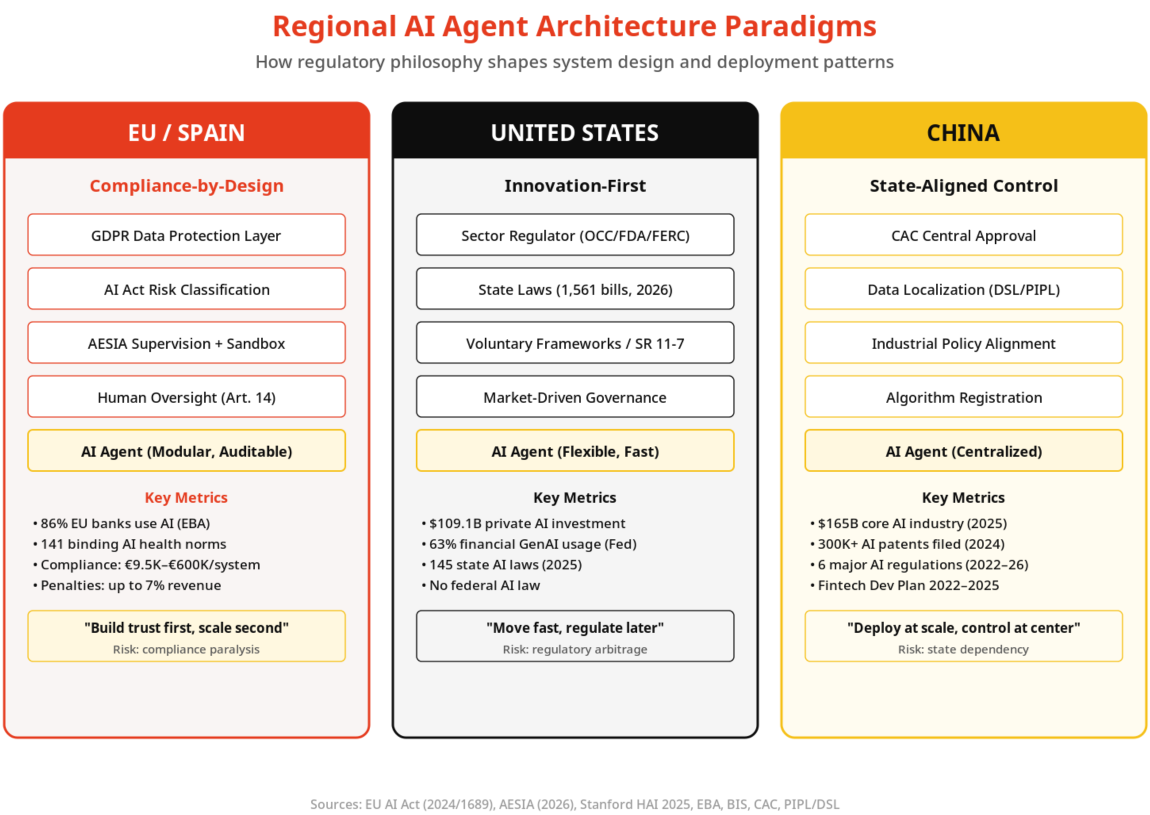

Figure 3. Regional AI Architecture Comparison — EU, US, China

RefleXio

www.reflexio.es
press@reflexio.es

RefleXio Intelligence

Independent intelligence on AI agents, regulation and execution in regulated sectors