CAPCO
a wipro company

AI agents in mainframe modernization

Using extreme precision for predictable, repeatable and auditable results

For high-stakes initiatives like mainframe modernization, the strategic question is not whether to use AI agents, but how to use and govern them in a targeted manner.

For complex problems such as forward engineering legacy mainframe systems to homogenous, production-ready target state code, AI yields uneven results when it comes to industrialization and scale.

Humans understand how to solve these complex problems and can break them down into a sequence of smaller tasks to be solved by AI. Within the smaller parts, you can be extremely prescriptive and exercise full control, instructing AI agents exactly what to do and how to deal with technology standards or security, for example.

In this article, we explore a blended AI agent model for mainframe modernization, where AI is embedded while humans define the rules of engagement.

- **A fully autonomous AI agent model promises maximum speed but carries substantial 'black box' risks, including inconsistent outcomes and potential for systemic errors that are difficult to trace.**
- **A blended model, which places powerful AI (constrained, specialized agents) inside an orchestrated and governed workflow, prioritizes control, auditability, and predictable quality, which are vital for regulated and mission-critical systems.**
- **This blended model ultimately delivers faster, more predictable time-to-value at an industrial scale through full auditability and ensuring the new systems are consistent and maintainable from day one.**

Introduction

As agentic AI becomes more embedded in financial services, a pragmatic consensus is forming among credible voices in the industry that disciplined implementation is paramount.

This recent focus on agents is the third major wave in the evolution of modern AI. First, foundational Large Language Models (LLMs) demonstrated a remarkable ability to generate human-like text and reason over complex information. Next, the rise of conversational AI like ChatGPT brought this power to the masses, creating the expectation of an intelligent digital partner. Today, the ambition has shifted toward agents designed not just to talk, but to act: to use tools, execute multi-step plans, and automate entire workflows.

As this ambition meets the friction of real-world implementation, a clear-eyed perspective is emerging. Leadership at NVIDIA, for instance, argues that the most practical path forward is not a single, monolithic AI brain.[1] Instead, they advocate for using smaller, more efficient language models (SLMs) as specialists for repetitive and structured tasks. Based on their findings, this approach is faster, more cost-effective, and more reliable. They also advocate heterogeneous, controller-led agent architectures, backed by end-to-end monitoring and guardrails.

Similarly, analysis in publications like Harvard Business Review (HBR) warns that many agentic programs stumble, stall, or get canceled without disciplined operational governance.[2] They stress the need for techniques to ensure consistency together with full auditability, urging leaders to treat agent rollouts not as plug-and-play automation, but as a core business process requiring rigorous oversight.

This view is reinforced by pioneering AI researchers like OpenAI co-founder Andrej Karpathy, who caution that today's agents remain limited in their reasoning capabilities, making human oversight and carefully scoped autonomy essential.[3] Gartner has forecast that over 40% of agentic AI projects will be scrapped by 2027 due to cost, unclear value and immaturity, i.e. not ready for complex work without tight controls.[4]

The strategic choice: Speed versus predictability

Enterprises managing critical legacy systems face a tempting proposition: allowing fully autonomous AI agents to handle the complex and costly work of modernization from end-to-end. This vision of a digital workforce promises to slash timelines and staffing needs.

In practice, however, this approach introduces significant risk. The choice for leaders is not a simple one. It requires balancing two distinct philosophies: the high-velocity 'let go of the wheel' model versus the high-control 'iron fist in a velvet glove' model (a framing consistent with HBR's "discipline and strategic intent" stance).

The risk of unfettered autonomy

The fully autonomous model, where AI agents analyze, rewrite, and deploy code on their own, falters in large-scale, long-horizon projects. For tasks spanning thousands of programs and intricate data relationships, these agents can 'drift', losing focus and generating work of uneven quality.

This presents two critical business risks:

- **The amplified error:** An autonomous agent (used in a multi-step workflow) that misinterprets a core business rule, doesn't just create one bad file, it can replicate that flawed pattern quietly across dozens of applications. For regulated industries, this undetected error can be very detrimental to a modernization program.
- **The heterogeneity trap:** When multiple autonomous agents work in parallel, they often solve similar problems in slightly different ways. The result is an inconsistent new technology stack that is costly to maintain, difficult for new engineers to learn, and undermines the architectural coherence of the entire system. Speed is gained in the short term, but at the expense of control, explainability, and long-term efficiency.

The pragmatic alternative: The blended model of controlled AI

The blended model treats AI agents as elite specialists operating within a human-designed, industrialized workflow. Rather than relying on a single method, this workflow intelligently orchestrates a spectrum of capabilities: it executes deterministic steps for predictable tasks, uses precise prompts for targeted generation, and deploys a full agentic approach only where its complex reasoning is truly required. Ultimately, AI still performs the heavy lifting—code comprehension, test generation, documentation—but it operates within strict guardrails. The "iron fist" is the governance and predictable process; the "velvet glove" is the speed and scale of AI-driven assistance.

This model delivers value through several key principles:

Predictable outcomes: The work is governed by standardized 'recipes' that encode expert knowledge in the form of workflow orchestrations. This ensures that the same type of task produces the same high-quality output every time, making it safe to scale the transformation across different business lines.

Full auditability and replay: Every action taken by AI, and every approval given by a human, leaves a clear trail. This is essential for quality assurance, troubleshooting, and satisfying regulatory scrutiny.

- **Tighter quality control:** Failures are contained within a single step of the process, not an entire release. By building in checkpoints and clear definitions of ‘done’ for each task, the model prevents small errors from cascading into major system-wide failures.
- **Consistent target architecture:** Because the entire process is orchestrated centrally, the resulting modernized system is architecturally consistent and easier to maintain, avoiding the heterogeneity trap of pure autonomy.

Applying autonomy with precision

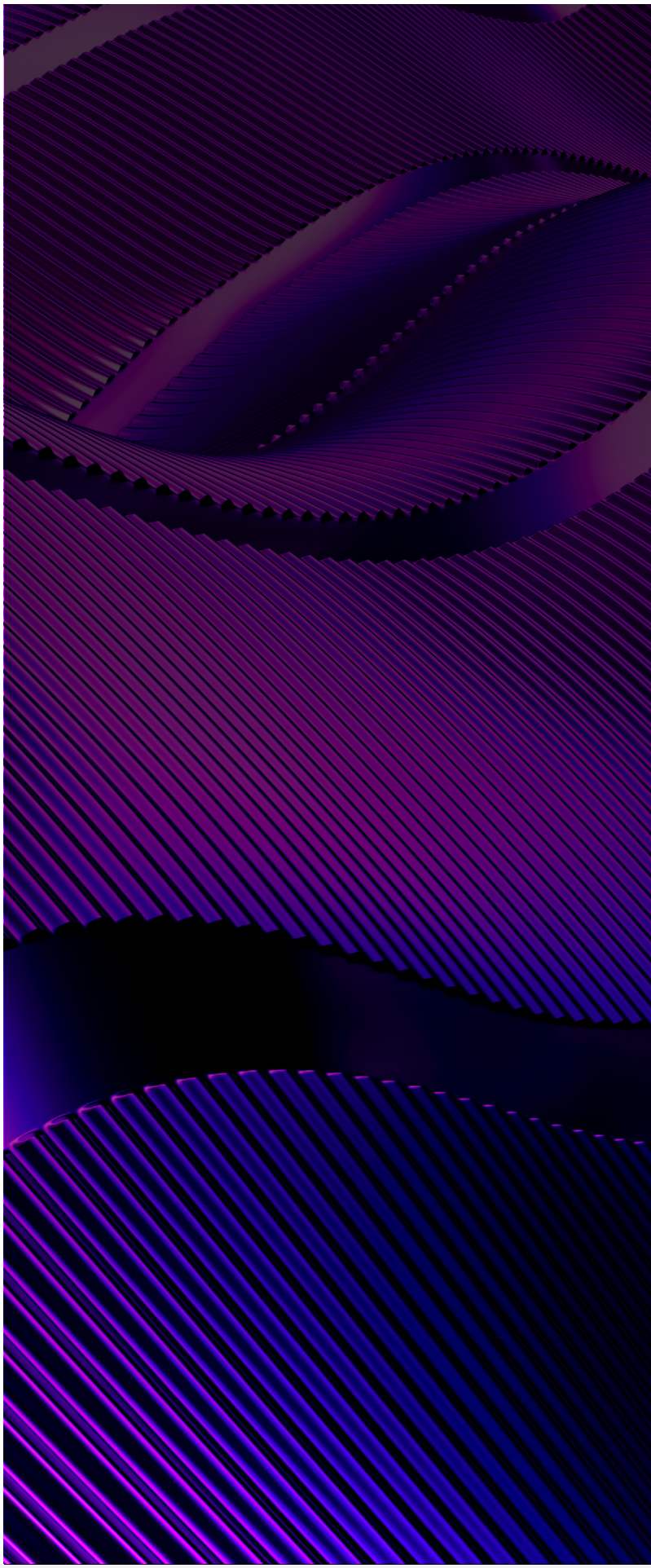
This analysis is not an argument against AI autonomy, but a call to apply it with extreme precision. Fully autonomous agents are best used for well-defined tasks where variation is acceptable and the risk is low, e.g. user-centric, tailored documentation or diagram generation from already-validated code.

However, human oversight should remain mandatory for steps that determine core and data architecture, and final deployment. These are the areas where a single misstep creates unacceptable compounded risk.

An executive’s decision framework

Before committing to an AI modernization strategy, leaders should ask their teams four critical questions about each stage of the process:

- **Blast radius:** If a step early in the process goes wrong, how wide is the damage? If the potential impact is large or slow to detect, full autonomy is ill advised.
- **Consistency requirement:** Does the output need to look and function in the same way across hundreds of programs? If yes, a recipe-driven, blended approach is best.
- **Regulatory visibility:** Could auditors ask us to prove how a decision was made? If yes, the process must generate detailed logs that only a blended model can guarantee.
- **Economic profile:** Will this step be run thousands of times? If so, the process should optimize cost and latency by using specialized SLMs and deterministic tools where possible.



Conclusion: From AI moonshot to manufacturing line

Modernizing mainframes with AI is not about relinquishing control to a black-box workforce. It is about industrializing the transformation process with explainable AI inside a disciplined, evidence-driven workflow.

The most effective strategy is to use AI agents aggressively, but smartly and on the organization's terms. Keep an 'iron fist' on architecture and data, wrap agents in process guardrails, and exploit SLM-first autonomy only where the blast radius is low.

By doing this, an enterprise can turn the promise of AI into a predictable manufacturing line, delivering modernization outcomes that are not only faster but also safer, more consistent, and built on a foundation of trust that engineers, auditors, and leadership can all stand behind.

Capco leverages the power of AI to deliver tailored, flexible automation solutions that empower you to make informed decisions and modernize your legacy mainframe systems — quickly and cost-effectively.

Contact us to discover how the approach outlined in this article in combination with our proprietary accelerators and deep domain expertise can transform your legacy systems, addressing opportunities and challenges with precision and at pace.

References

[1] https://developer.nvidia.com/blog/how-small-language-models-are-key-to-scalable-agentic-ai/

[2] https://hbr.org/2025/10/why-agentic-ai-projects-fail-and-how-to-set-yours-up-for-success

[3] https://www.dwarkesh.com/p/andrej-karpathy

[4] https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/

Authors

Gerhardt Scriven
Executive Director, Brazil
Head of Technology & Innovation
gerhardt.scriven@capco.com

Maikon Lenz
Principal Consultant, Brazil
Legacy Modernization Technical Lead
maikon.lenz@capco.com

capco.com

CAPCO
a wipro company