finshape

From Pilots to Agents:

How Banks Move Beyond GenAI Experiments to an Agentic Operating Model

Josef Dvořák
AI Director Finshape

Executive Summary

AI is transforming financial services, yet returns often lag behind expectations

Around 65–72 % of organizations already use AI, and about 65% have deployed generative AI in at least one area[1]. Still, only 1% have scaled it successfully and achieved real financial impact[2].

Most pilots stall because **AI is added onto old workflows instead of reshaping products and processes** around it. Regulation raises the bar further: the EU AI Act and the Digital Operational Resilience Act (DORA) demand post-market monitoring, incident reporting, and human oversight for high-risk systems [3].
To meet these requirements, **leading banks are forming AgentOps teams** responsible for observability, guardrails, and incident response, often running "old + new" operations side by side [4].

When done right, the benefits AI brings are clear:

A major Czech bank saved the equivalent of **30 FTEs** in its call center through AI-driven voice automation.

HSBC doubled to quadrupled true-positive AML detections and cut false positives by 60%, shortening investigations from weeks to days [5].

Swedbank's "Ida" assistant now manages over 2 million interactions a year, resolving about 80% of cases on the first contact [6].

Morgan Stanley Wealth Management reports near-universal adoption of its AI assistant across advisory teams, driven by high-quality, well-integrated knowledge workflows [7].

NatWest's "Cora+" upgrade improved customer satisfaction by 150% and reduced human involvement [8].

BNY Mellon now employs "digital employees" who perform supervised back-office tasks like coding and payment validation [9].

These examples confirm that AI pays off, but only when banks redesign work, embed oversight, and follow a phased roadmap – from pilots to controlled autonomy.

Chapter 1

From hype to hard numbers: Why pilots stall

AI is everywhere in banking – but measurable impact is rare.

While most institutions now use some form of AI, only about 1% operate it at scale and can prove sustained financial returns [2]. Many projects never make it past the pilot phase.

The main barriers are structural, not technical. They include:

- **Fragmented data and controls.** Pilots often rely on isolated datasets without monitoring or logging, blocking compliance with **post-market obligations** [3].
- **No dual-run operations.** Without "shadow mode" – where new AI flows run in parallel to legacy ones – banks can't compare metrics or build confidence.
- **No AgentOps function.** Few teams actively monitor agent behavior, evaluate outputs, or manage incidents [4]. Once deployed, systems go dark.
- **Unchanged human roles**. Advisors and analysts still work from manual playbooks. Without updated responsibilities or incentives, AI sits unused.
- **Regulatory uncertainty.** Compliance teams fear breaching consumer-protection or AML rules, especially as the **AI Act** demands auditable logging, risk management, and **human oversight** [3].

IN SHORT:

Technology alone doesn't deliver ROI. True transformation requires redesigning workflows, defining human–AI collaboration, and embedding operational oversight.

Chapter 2

What is an “agentic bank”?

The next stage of AI in banking goes far beyond chatbots.

It’s about autonomous agents – systems that can plan, decide, and act under human oversight.

They don’t just answer questions; they complete tasks – processing payments, updating customer data, or resolving complaints.

Key elements of an agentic bank:

- **Task-oriented agents.** These AI systems handle multi-step tasks end-to-end – for example, resetting passwords, resolving disputes or automating KYC remediation.
- **Human-in-the-loop.** Agents operate within clear thresholds and seek human approval for sensitive or high-risk actions. Oversight ratios and escalation playbooks ensure control and accountability.
- **AgentOps.** Think of it as mission control for AI. A dedicated operations layer that tracks agent activity, logs interactions, evaluates outputs, and triggers rollbacks if rules or safety limits are breached [4].
- **Dual-mode operations.** Legacy and AI-driven processes run in parallel until AI systems prove their value and reliability.

Adopting this model also means accepting that retrieval – augmented generation (RAG) – even when grounded in internal data – still hallucinates about 17% of the time [11]. That’s why banks need continuous monitoring for accuracy and alignment before letting AI take real actions.

Ultimately, an “agentic bank” is one where AI acts, humans decide – and oversight never sleeps.

“The goal isn’t to replace people with agents – it’s to give every employee a digital teammate who never gets tired.”

Petr Štěpánek, *Head of Applied AI, Finshape*

Chapter 3

Evidence that it works when governed

When banks pair AI with strong governance and human oversight, results scale safely and fast.

HSBC – fighting financial crime
Processes **900 million transactions monthly** across 40 million accounts. AI now detects **2–4× more** true positives and **60% fewer false alerts**, cutting investigations from weeks to days [5].

Swedbank – customer service at scale
Virtual assistant Ida handles **2 million** + conversations yearly with **~80% first-contact resolution** [6]. Clear guardrails and human hand-offs ensure trust.

Morgan Stanley – advisor copilot
Nearly all advisor teams use its internal AI assistant daily – proof that r**eliable content + workflow fit = adoption** [7].

NatWest – Cora+ upgrade
Generative AI lifted customer satisfaction by 150% and cut the need for human support [8].

BNY Mellon – digital employees
Supervised AI "employees" now handle coding and payment validation under strict audit controls [9].

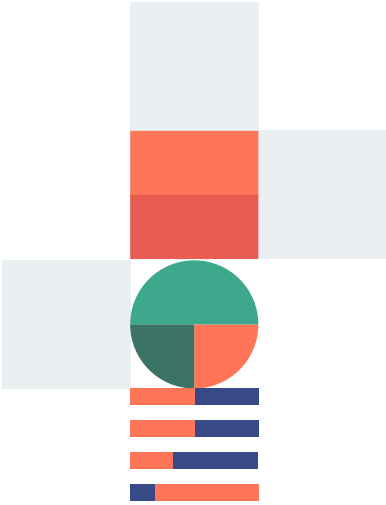
These examples show that AI succeeds when technology, governance, and people align.

Chapter 4

Architecture & controls: How to operate agents safely

Building an "agentic bank" isn't only about AI capabilities – it's about control, visibility, and safety by design.

A solid architecture allows agents to act with confidence and stay within regulatory and operational boundaries.



Reference architecture

- **Channels and surfaces.**
  Agents should **work consistently across all touchpoints** – web, mobile, call centers, and advisor desktops – under a unified policy framework.

- **Orchestration layer (AgentOS).**
  This is the brain of the system. It manages planning, tool use, and memory, applying **role-based permissions, rate limits, and kill switches** that can instantly pause an agent when thresholds are breached [4].

- **Retrieval and knowledge layer.**
  Agents use policy-aware RAG systems that pull from verified sources with freshness and attribution checks. Even with RAG, groundedness detectors are essential – studies show **AI tools still hallucinate about 17% of the time** [11].

- **Safety and compliance.**
  Content moderation, PII protection and toxicity filters are mandatory. Banks must also run **AI-safety evaluations** to test groundedness, factuality, and policy alignment.

- **Observability (AgentOps).**
  Continuous monitoring tracks real-time metrics such as groundedness, tool errors, and policy violations. Event streams are recorded for replay, escalation, or rollback. Regular **red-team drills** help validate kill switches and incident alerts [4].

Regulatory alignment

- **Digital Operational Resilience Act (DORA).**
  Effective since **January 2025**, DORA mandates ICT-risk management, incident classification, resilience testing, and third-party oversight for all financial entities – including those using AI [3].

- **EU AI Act.**
  Fully applicable from **August 2026**, it requires **post-market monitoring, human oversight**, and **incident reporting** for high-risk systems [3].

- **NIST AI Risk Management Framework.**
  Offers a global standard for documenting and auditing AI systems, helping banks align with both European and international expectations.

By combining AgentOS orchestration, AgentOps observability, and regulatory compliance, banks can safely shift from pilots to production without losing control of outcomes.

"AI freedom requires discipline. The best systems aren't those that do everything – but those that know when to stop."

Sára Hanniker, *Head of AI core and solutions, Finshape*

Chapter 5

KPI tree & business case

Scaling AI requires more than enthusiasm – it needs proof.

A clear KPI framework links technology outcomes to business value and regulatory expectations.

Service & sales

- **Deflection vs. containment.** Share of customer requests handled without humans. Top performers achieve **60–80% first-contact resolution (Swedbank ~80%)** [6].
- **Average handle time (AHT).** AI-assisted conversations should be **10–30 % faster**; NatWest saw major gains after deploying GenAI [8].
- **Conversion & cross-sell.** Measurable uplift when advisers use copilots; Morgan Stanley achieved near-universal adoption and productivity growth [7].

Financial crime

- **True vs. false positives.** Tracking model accuracy drives ROI. **HSBC** reported **2–4× more true positives and 60% fewer false alerts** [5].
- **Investigation cycle time.** AI shortened case closure from weeks to days, directly lowering compliance costs [5].

Operations & finance

- **Autonomous resolution rate.** Measures tasks handled without human input; BNY Mellon's digital employees already perform supervised production work [9].
- **Mean time to repair.** Faster recovery after process interruptions indicates operational maturity.

Trust & safety

- **Groundedness / factuality scores.** Even RAG systems still hallucinate about 17% of the time, so human review for unverified outputs is mandatory [11].
- **Policy-violation rate & incident escalations.** Metrics must feed back into AgentOps monitoring and regulatory reporting.

These KPIs tie directly to AI Act and DORA obligations – tracking incidents, monitoring performance, and ensuring human oversight. When boards see clear metrics linking AI to efficiency, risk reduction, and trust, investment confidence follows.

"What gets measured gets scaled. Without the right KPIs, AI stays a science experiment."

Michal Kalousek, *Business Consulting Director, Finshape*

Chapter 6

The roadmap:
From pilot to agentic bank in 2 years

Transforming from scattered AI pilots to a fully agentic operating model doesn't happen overnight.

It requires a phased, disciplined approach – balancing innovation with governance and scaling only when measurable readiness is achieved.

Below is a two-year roadmap that has emerged from early adopters' best practices.

Phase 0 (weeks 0–4) Stabilize & govern

- Establish an **AI risk committee** owned jointly by CIO and COO.
- Map **DORA** controls to AI operations and adopt the **NIST AI Risk Management Framework** for documentation [3].
- Take inventory of all existing AI use cases, data sources, and third-party dependencies.
- Define early **metrics and escalation rules** – every experiment must be observable from day one.

Goal: create transparency and governance before scale begins.

Phase 1 (months 1–6) Shadow-mode pilots

- Select **2–3 high-value processes** such as service resolution, AML triage, or reconciliation.
- Run agents in read-only **"shadow mode"**, where humans make final decisions but AI provides recommendations.
- Collect **baseline metrics** – accuracy, handle time, false-positive rates – to compare old vs. new.
- Implement trace logging, groundedness scoring, kill switches and escalation paths [4].

Goal: prove value and build confidence without operational risk.

Phase 2 (months 7–12) Copilot to controlled autonomy

- Gradually allow **auto-approval** for narrow tasks with low risk (defined thresholds for amount, customer type, or error margin).
- Keep humans in approval loops for all other actions.
- Launch **post-market monitoring dashboards and incident-response drills** to test escalation speed and kill-switch reliability [4].
- Introduce continuous evaluation and fine-tuning cycles based on feedback.

Goal: move from observation to measurable automation under tight oversight.

Phase 3 (months 13–24) Scale & redesign services

- Form a permanent **AgentOps team** that consolidates monitoring, policy management, and red-teaming [4].
- Redesign customer channels for **task completion**, not only conversation – e.g., secure document upload, appointment booking, or transaction execution.
- Integrate AI agents into **advisor and employee desktops** to augment productivity.
- Expand across domains such as **wealth management, SME lending, and customer onboarding** once safety and KPIs prove maturity.

Goal: achieve an enterprise-wide model where human and AI collaboration becomes standard, measurable, and compliant.

Ultimately, this roadmap helps banks transition from pilots to production responsibly – scaling what works while maintaining full control.

"The real race isn't who launches AI first – it's who scales it safely."

Sára Hanniker, *Head of AI core and solutions, Finshape*

Chapter 7

Operating model & roles

Agentic banking isn't just new tech – it's a new way of working.

Clear roles keep AI safe, effective, and compliant.

Key roles in an agentic model:

Product Owners: define agent scope, KPIs, and approval limits.

Business Designers: redesign workflows and map where AI adds value.

AgentOps/SRE Engineers: monitor performance, evaluate behavior, and run drills [4]; ensure logging and post-market checks required by the AI Act [3].

Risk & Compliance: align with AI Act and DORA, manage incident reporting and vendor risk [3].

Data & Platform Engineers: maintain retrieval pipelines, red-team tests, and safety filters.

Front-line Staff: manage escalations, validate outputs, and feed lessons back to improve workflows.

Together they form the backbone of a resilient, continuously learning AI organization.

Chapter 8

Build vs. buy: Decision guardrails

Every bank faces a key decision: develop its own agentic platform or assemble it from vendors. The right answer depends on control, speed, and compliance priorities.

Build when

- **Deep integration** with internal systems and data sovereignty is essential.
- You need custom orchestration, bespoke guardrails, or sensitive workflows.
- **Long-term flexibility and IP ownership** outweigh time-to-market pressures.

Buy or assemble when

- **Rapid deployment** and packaged compliance features matter more than full control.
- Vendors **can deliver** built-in observability, groundedness scoring, and policy engines for approvals and entitlements [3].
- Modular, portable architectures prevent vendor lock-in while **meeting DORA and AI Act requirements.**

Owning the brain may be strategic, but renting the muscles keeps you fast.

Chapter 9

Questions board will ask (and how to answer)

Boards demand clear evidence that AI initiatives are safe, valuable, and compliant.

Anticipating their questions – and answering them with facts – builds credibility and ccelerates scale.

1. **Where's the ROI?**
   Show concrete examples:
   - → **HSBC** cut false positives by **60%** and doubled-to-quadrupled true AML detections [5].
   - → **Swedbank** achieves **~80% first-contact resolution** [6].
   - → **Morgan Stanley** reports near-total adoption of its AI copilot [7].
   - → **NatWest** improved customer satisfaction by **150%** [8].

   Link savings to avoided fines, lower headcount growth, or revenue lift – not just "efficiency."

2. **Is it safe?**
   Demonstrate **post-market monitoring, human-in-the-loop thresholds, and kill switches** that can instantly stop rogue behavior. Map all controls to the **AI Act** (Article 72) and **DORA** incident processes [3][4][11].

3. **How will we staff it?**
   Present the **AgentOps model**: dedicated monitoring, SRE-style on-call rotations, and upskilling for advisers and analysts [4].

4. **When can we scale?**
   Outline clear value gates – measurable KPIs such as first-contact resolution, handle time, or false-positive reduction – that must be achieved before broader rollout.

Boards invest with confidence when AI is framed not as an experiment, but as a controlled, auditable operating model.

Chapter 10

Finshape angle: How we can help

Finshape is evolving its platform into an Agentic Digital Banking Operating System (DBOS) – a low-code environment powered by an AI Expert Network.

This model turns Finshape's deep banking expertise into a coordinated AI workforce that helps banks accelerate transformation safely and efficiently.

Core components:

- **AI expert network.** Specialized agents – domain experts, analysts, architects, testers, and deployment managers – collaborate to turn requirements and Figma designs into complete configurations, test suites, and deployment scripts with human-in-the-loop safety mechanisms.
- **Knowledge layer.** Agents draw on Finshape's documentation, APIs, compliance rules, and trusted third-party sources to ensure accuracy and regulatory alignment.
- **Orchestration engine.** A multi-agent coordination framework manages planning, scheduling, and monitoring, enforcing communication protocols and safety constraints.
- **Human collaboration points.** Validation, design approval, and deployment authorization are built in, ensuring humans stay in control.
- **Embedded observability & compliance.** Full logging, dashboards, and metrics align with DORA and the AI Act, including kill switches and approval workflows.

Partnering with Finshape gives banks access to a ready-made digital workforce and the framework to deploy AI-driven services safely and at scale.

"Finshape doesn't just bring tools – it brings an AI workforce that knows banking."

Petr Koutný, *CEO, Finshape*

Turning insight into action

We publish our books to inspire, share know-how, and open conversations about the future of banking and innovation.

Who we are

finshape

We build the future of banking. With over 30 years of experience, more than 500 professionals, and a presence in 25+ countries, Finshape delivers a modular, cloud-native Digital Bank Operating System (DBOS) that gives banks digital sovereignty, speed, and scalability.

Recent moves – such as the acquisition of a loyalty platform – deepen our ability to make banking more personalized, engaging and profitable.

Direct People
A FINSHAPE COMPANY

Innovation is in our DNA. Since 2010, Direct People has been crafting human-centered products that succeed in markets.

With over 60 launches, 300+ clients across 11 industries, and a deep expertise in design, AI, and technology, we bring ideas to life. Now, as a Finshape company, we bridge strategy, user insight and delivery into one full-stack force.

Ready to take the next step?
Contact the expert.

Josef Dvořák
AI Director Finshape
+420 728 286 466
josef.dvorak@finshape.com

Appendix

References

- **McKinsey & Company, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value" (2024).** The report notes that about 65 % of respondents' organisations regularly use generative AI and 72 % use AI in at least one business function [1][2]. → *www.mckinsey.com*

- **Digital Marketing Institute, "Is your organisation ready for AI?" (2024) and Fortune / MIT report summary (2024).** Digital Marketing Institute highlights that o**nly about 1 % of organisations that have adopted generative AI consider their investments to be mature [3].** A Fortune article summarising an MIT report notes that only about 5 % of gen-AI pilots generate rapid revenue acceleration, while the vast majority stall [4]. → *digitalmarketinginstitute.com*

- **European Commission, "AI Act – Regulatory framework for AI" (updated 1 August 2025) and European Securities and Markets Authority (ESMA), "Digital Operational Resilience Act (DORA)" (2025).** The AI Act entered into force **on 1 August 2024 and will be fully applicable on 2 August 2026**, with some obligations applying earlier; it requires high-risk AI systems to implement post-market monitoring, logging and human oversight [5]. ESMA notes that DORA entered into force **on 16 January 2023 and applies as of 17 January 2025**, imposing ICT-risk management, incident reporting and resilience testing [6]. → *digital-strategy.ec.europa.eu*

- **XenonStack, "AgentOps: the next evolution in AI lifecycle management" (2024).** The article explains that AgentOps extends MLOps and LLMOps and that dynamic AI agents require reliability, monitoring of decision-making processes, debugging of multi-step interactions and end-to-end observability[7]. → *www.xenonstack.com*

- **Google Cloud press release, "Google Cloud launches AI-powered anti-money-laundering product for financial institutions" (21 June 2023), and HSBC, "Harnessing the power of AI to fight financial crime" (2025).** The press release states that HSBC's AML AI detects **2–4× more true positives, reduces false positives by 60 % and cuts batch processing from several weeks to a few days** [8][9]. → www.googlecloudpresscorner.com

- HSBC reports that it monitors **about 900 million transactions each month across 40 million customer accounts** and uses AI to detect financial crime [10]. → www.hsbc.com

- **Nuance case study, "Swedbank delivers fast, convenient self-service with Nuance" (2022) and Diginomica, "Swedbank puts customers on track for self-service satisfaction" (2016).** The Nuance case study notes that Swedbank's virtual assistant handles **more than two million conversations a year and answers around 80 % of questions during customers' initial engagement** [11][12]. → www.nuance.com/content

- The Diginomica article reports that the assistant handled **around 30,000 contacts per month with a first-contact resolution rate pushing 80 %** during its initial deployment[13]. → https://diginomica.com

- **OpenAI, "Morgan Stanley uses AI evals to shape the future of financial services" (2024).** The article notes that **over 98 % of Morgan Stanley's advisor teams actively use the internal AI assistant** [14]. → www.finextra.com

- **Finextra, "BNY to give AI-powered digital employees their own emails" (July 2025)**. Finextra reports that Bank of New York Mellon has dozens of AI-powered "digital employees" with their own logins and managers who perform tasks such as code clean-up and payment-instruction validation; the bank plans to give them email accounts [16]. → www.finextra.com

- **Stanford Law School & Thomson Reuters Institute, "Hallucination-Free? Assessing the reliability of leading AI legal research tools" (2025).** The study finds that retrieval-augmented AI tools hallucinate between **17 % and 33 % of the time** [17], underscoring the need for groundedness detectors. → dho.stanford.edu