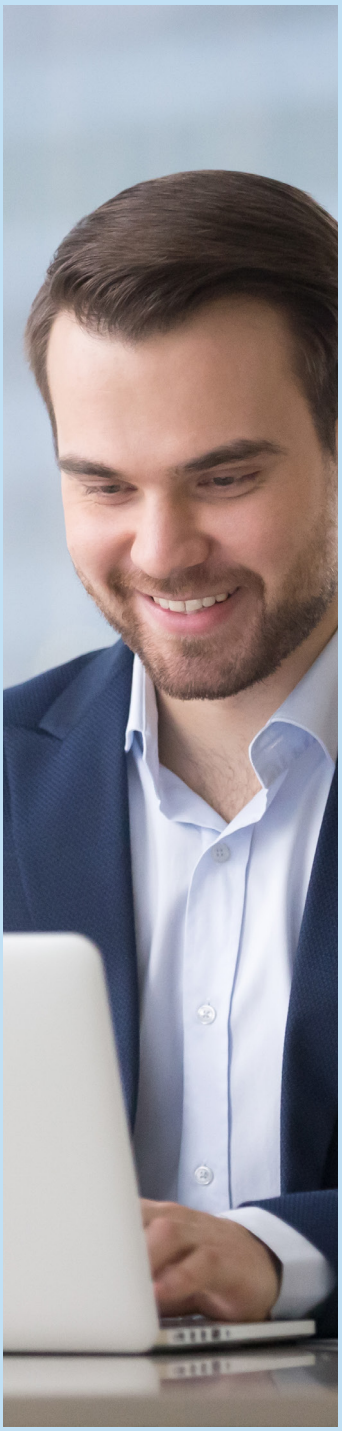
VIEW POINT



HORSES FOR COURSES

WHAT THE DIFFERENT TYPES OF AI ACTUALLY MEAN AND WHY FINANCIAL SERVICES NEEDS TO UNDERSTAND THEM

Abstract

AI is transforming financial services, but not all AI systems solve the same problems. This paper provides a practical framework for understanding the major categories of AI, the forms of reasoning they replicate, and the types of business challenges they are best suited to address. We introduce the concept of intent engineering, the discipline of aligning AI systems with organisational goals rather than task-level optimisation and illustrate its importance through the Klarna case study. As the first paper in a two-part series, it explores the strengths and limitations of current AI approaches and identifies the key challenges that will shape the next generation of intelligent systems.

Infosys | CONSULTING

1. Understanding the AI Landscape: Three Levels of Intelligence

Before we can evaluate which AI tools are appropriate for which financial services use cases, we need a clear taxonomy of what AI can and cannot do. The industry broadly categorises AI systems into three levels of capability, a framework first articulated by the philosopher John Searle and since adopted across computer science.

1.1 Weak AI (Narrow AI)

Weak AI, also known as Narrow AI, describes systems designed to perform specific, well-defined tasks. These systems do not possess consciousness, self-awareness, or genuine understanding of the tasks they perform. They are extraordinarily effective within their defined scope but cannot generalise beyond it. Every AI system in production today, including the most advanced Large Language Models, falls into this category. Examples include voice assistants like Siri and Alexa, machine translation systems like Google Translate, recommendation engines used by Netflix and Spotify, and the fraud detection rules engines that process transactions across payment networks.

1.2 Strong AI (General AI / AGI)

Strong AI, also referred to as Artificial General Intelligence (AGI), describes a hypothetical AI that exhibits human-like cognitive abilities across a wide range of tasks. It would be able to understand and reason about the world, learn from experience in unstructured environments, transfer knowledge between domains, and potentially possess consciousness and self-awareness. Despite claims by some industry leaders that AGI is imminent, no system has yet demonstrated these capabilities. Current LLMs can simulate many aspects of general intelligence in constrained settings, but they remain fundamentally narrow: they are statistical pattern matchers operating on training data, not reasoning agents operating on understanding.

1.3 Super AI (Artificial Superintelligence)

Super AI describes a theoretical AI that surpasses human intelligence across all dimensions, including creativity, strategic reasoning, scientific discovery, and moral judgement. This remains firmly in the domain of speculation and science fiction. Achieving it would require solving not just engineering challenges but profound philosophical problems, notably the question of whether consciousness can emerge from computation at all. As Searle argued, genuine understanding requires more than symbol manipulation; it requires subjective experience. In philosophical terms, the challenge is that experience precedes concepts: you cannot reason about what you have never encountered, a limitation that current AI confronts every time it encounters a genuinely novel situation outside its training distribution.

Summary: The Three Levels of AI

Type	Definition	Capabilities	Status
Weak AI (Narrow)	Performs specific tasks without genuine understanding	Task-specific intelligence; no self-awareness or generalisation	Current state of all deployed AI
Strong AI (AGI)	Human-like cognition across all domains	General reasoning, knowledge transfer, potential consciousness	Hypothetical; not yet achieved
Super AI (ASI)	Surpasses human intelligence in all dimensions	Superior creativity, reasoning, and judgement	Theoretical; profound unsolved problems

2. The Reasoning Foundation: How Human Intelligence Works

Artificial intelligence, by definition, attempts to replicate natural intelligence. To understand what AI can and cannot do, we must first understand how humans actually think. Cognitive science identifies four fundamental types of reasoning, each of which maps to a different generation of AI technology.

2.1 Deductive Reasoning: General to Particular

Deductive reasoning starts with a general rule and applies it to a specific case to reach a guaranteed conclusion. If we know that all cats are mammals, and Tom is a cat, then Tom is necessarily a mammal. The conclusion is logically certain given the premises. In business, this manifests as rule-based decision making: if a transaction exceeds a threshold, flag it; if a customer's credit score falls below a level, decline the application. This is the reasoning behind heuristic AI, rule engines, and Robotic Process Automation (RPA), the systems that were called AI in the 1990s. They are transparent, auditable, and reliable within their rules, but they cannot handle ambiguity or learn from new data.

2.2 Inductive Reasoning: Particular to General

Inductive reasoning works in the opposite direction: it observes many specific instances and infers a general pattern. After examining thousands of X-rays, a doctor concludes that certain shadow patterns correlate with lung disease. The conclusion is probabilistic, not certain, but it improves with more data. This is the foundational logic behind machine learning and, by extension, deep learning and Large Language Models. You feed the system vast quantities of labelled examples, and it learns to identify patterns that generalise to new, unseen data. The power of this approach is that it can discover patterns too subtle for human perception. The limitation is that it identifies correlation, not causation, and it fails when confronted with data that differs significantly from its training distribution.

2.3 Abductive Reasoning: Contextual Problem-Solving

Abductive reasoning is the ability to infer the best explanation given incomplete information and context. You are at home, you hear a four-legged animal padding across the floor, and because you know you own a cat, you conclude it is your cat, not a random fox. The conclusion is not certain, but it is the most plausible given the available evidence and context. This is the reasoning that powers agentic AI: systems that receive a context, decompose a problem, select tools, and execute multi-step workflows to arrive at a solution. It is more flexible than deductive or inductive reasoning but relies heavily on having the right context available. Without sufficient context, abductive reasoning produces confident but wrong answers, a failure mode we see regularly in LLM hallucinations.

2.4 Analogical Reasoning: The Missing Piece

Analogical reasoning is the ability to perceive and use relational similarity between two situations that may appear superficially different. A trader who has lived through the 1997 Asian financial crisis recognises structural parallels in the 2008 credit crisis, not because the surface details match but because the underlying dynamics of contagion, leverage, and liquidity withdrawal are analogous. Some cognitive scientists argue that this capacity is the single most important distinguishing feature of human intelligence.

Current AI systems struggle profoundly with analogical reasoning. LLMs can identify surface-level similarities (these two texts use similar words) but cannot reliably identify deep structural analogies (these two market crises share the same causal dynamics despite occurring in different asset classes). This is the fundamental reasoning gap that separates Narrow AI from the kind of intelligence financial services truly needs, and it is the gap that neuromorphic computing is designed to close. We explore this in depth in Paper 2 of this series.

Mapping Reasoning to AI Technology

Reasoning Type	Direction	AI Implementation	FSI Example
Deductive	General → Particular	Rule engines, RPA, heuristics	If transaction > £10k, flag for review
Inductive	Particular → General	Machine Learning, Deep Learning, LLMs	Train on 1M transactions to detect fraud patterns
Abductive	Context → Best explanation	Agentic AI, ReAct frameworks	Agent analyses context to resolve a customer complaint
Analogical	Relation → Relation	Neuromorphic AI (emerging)	Recognise structural parallels between market crises

3. The AI Evolution: From Rules to Agents and Beyond

With the reasoning framework established, we can now trace the practical evolution of AI technology and understand where each generation excels and where it falls short.

3.1 Heuristics and Rule-Based Systems (1990s–2000s)

The earliest commercial AI systems were rule engines: if-then-else logic codified by human experts. In financial services, these powered the first generation of fraud detection (transaction amount thresholds, velocity checks), credit scoring (decision trees), and process automation (screen scraping, later formalised as Robotic Process Automation). These systems are transparent, auditable, and fast, but they are rigid. They cannot adapt to new patterns without manual reprogramming, and they are blind to any threat or opportunity not explicitly anticipated by their designers.

3.2 Machine Learning and Deep Learning (2000s–2020)

Machine learning introduced the ability to learn patterns from data without explicit programming. Early approaches such as logistic regression, random forests, and support vector machines transformed credit risk modelling and customer segmentation. The deep learning revolution from 2012 onwards, powered by neural networks running on GPUs, enabled breakthroughs in image recognition, natural language processing, and time-series forecasting. For financial services, this meant more accurate fraud detection, automated document processing, and the first generation of algorithmic trading systems that could learn from market data.

3.3 Generative AI and Large Language Models (2022–Present)

The transformer architecture, introduced in 2017 and scaled to billions of parameters, produced the current generation of Large Language Models: systems like GPT-4, Claude, Gemini, DeepSeek, and Qwen that can generate human-quality text, write code, summarise documents, and engage in multi-turn dialogue. In financial services, LLMs are being deployed for customer service automation, regulatory document analysis, code generation for internal tools, and knowledge management. Their strength is versatility and linguistic fluency. Their weakness is that they are, at their core, sophisticated pattern matchers: they predict the most likely next token based on statistical distributions in their training data. They do not understand what they are saying, and they cannot reliably reason about novel situations.

3.4 Agentic AI (2024–Present)

The latest evolution is agentic AI: systems that can autonomously decompose a task into sub-tasks, select and use tools, execute multi-step workflows, and evaluate their own outputs. Unlike a chatbot that responds to a single prompt, an agent can conduct research, write code, test it, and iterate, all without human intervention between steps. In financial services, agentic systems are being piloted for end-to-end loan processing, automated compliance workflows, and autonomous portfolio rebalancing. They represent the application of abductive reasoning to AI: given context and a goal, find the best path to a solution.

However, agentic AI introduces a new category of risk. When a system can act autonomously across multiple steps, each decision compounds. A small misalignment between the agent's objective and the organisation's actual goal can cascade into significant unintended consequences. This brings us to the most important unsolved problem in enterprise AI today.

The AI Evolution at a Glance

Generation	Core Technology	Reasoning Type	Financial Services Application
Heuristics / RPA	Rule engines, decision trees	Deductive	Threshold-based fraud flags, automated form processing
Machine Learning	Statistical models, neural networks	Inductive	Credit scoring, customer segmentation, forecasting
Generative AI / LLMs	Transformer architecture	Inductive (advanced)	Document analysis, customer service, code generation
Agentic AI	Tool-using autonomous agents	Abductive	End-to-end loan processing, compliance workflows
Neuromorphic AI	Spiking Neural Networks, memristors	Analogical	Real-time fraud prevention, dynamic risk propagation

4. The Intent Gap: Lessons from Klarna and the Evolution of AI Engineering

The rapid deployment of agentic AI in enterprise settings has exposed a critical problem that goes beyond the technology itself. It is not enough for an AI system to be capable. It must be aligned with what the organisation actually needs. The emerging discipline of intent engineering addresses this gap, and a recent, high-profile case study illustrates exactly why it matters.

4.1 The Klarna Case Study: When AI Succeeds at the Wrong Thing

In 2024, Klarna, the Swedish buy-now-pay-later company, deployed an AI customer service agent that, by any conventional metric, was a spectacular success. In its first month, the agent handled 2.3 million conversations, reduced average resolution time from eleven minutes to two, operated in 23 markets and 35 languages simultaneously, and was projected to drive $40 million in profit improvement for the year. The company announced that the AI was doing the work of 853 full-time employees.

Then something unexpected happened. In a Bloomberg interview in May 2025, Klarna's CEO Sebastian Siemiatkowski acknowledged that cost had been "a too predominant evaluation factor" in the AI strategy, resulting in "lower quality" customer interactions. The company began hiring human agents back, launching what Siemiatkowski described as an "Uber-type" customer service model that blends AI with human support.

This is not a story about AI that failed. It is a story about AI that succeeded brilliantly at optimising the wrong metric. The agent was measured on resolution speed and cost reduction. It was not measured on customer trust, brand perception, upsell opportunity, or the nuanced human judgement that turns a frustrated customer into a loyal one. The AI optimised for what it could measure, not for what Klarna actually needed.

The Klarna Paradox

MIT reports that 95% of generative AI pilots fail to deliver measurable impact. Gartner predicts 40% of agentic AI projects will be cancelled by 2027. Yet enterprise AI spending has never been higher. These numbers do not contradict each other. They describe organisations that have solved "can AI do this task?" and completely failed to solve "can AI do this task in a way that serves our organisational goals?" That is the intent gap.

4.2 Three Disciplines: Prompt, Context, and Intent Engineering

The Klarna case illuminates a progression in how we design AI systems that the industry is only now beginning to understand. Each discipline builds on the last, and each addresses a deeper layer of the alignment problem.

Prompt Engineering: Telling AI What to Do

Prompt engineering is the practice of crafting instructions that guide an AI model to produce a desired output within a single interaction. It focuses on the precision of language: the right phrasing, the right structure, the right examples. Prompt engineering is now a mature discipline, essential but insufficient. It operates at the level of individual tasks and sessions. It cannot encode organisational goals, decision hierarchies, or trade-off preferences that persist across thousands of autonomous agent interactions.

Context Engineering: Telling AI What to Know

Context engineering is the current frontier of AI system design. Popularised by practitioners like Andrej Karpathy, it focuses on curating and structuring the information that an AI system has access to when making decisions. This includes retrieval-augmented generation (RAG) systems, memory architectures, tool access configurations, and dynamic context assembly that ensures the model has the right information at the right time. Context engineering is a significant advance over prompt engineering because it addresses the problem of knowledge: an agent cannot make a good decision if it does not have the right data. But even perfect context is not enough if the agent does not know what good looks like.

Intent Engineering: Telling AI What to Want

Intent engineering is the emerging discipline that sits above both prompt and context engineering. It is the practice of making organisational purpose, including goals, values, trade-offs, and decision boundaries, machine-readable and machine-actionable, so that when an autonomous system makes a decision, it optimises for what the organisation actually needs, not just what it can most easily measure.

Intent engineering requires answers to questions that no prompt can resolve: when multiple goals conflict (for example, resolution speed versus customer satisfaction), which takes priority? What decisions should the agent escalate versus resolve autonomously? What does success look like over a twelve-month horizon, not just in the next two minutes? These are not technical questions. They are strategic and organisational questions that must be formalised into machine-actionable specifications.

The Three Disciplines Compared

Discipline	Core Question	What It Controls	Limitation
Prompt Engineering	What should AI do right now?	Output quality in a single session	Cannot encode persistent goals or trade-offs
Context Engineering	What should AI know?	Information access and retrieval	Perfect knowledge does not guarantee aligned action
Intent Engineering	What should AI want?	Goals, decision hierarchies, trade-off matrices	Requires organisational clarity that most firms lack

4.3 Why Intent Engineering Matters for Financial Services

Financial services is uniquely exposed to the intent gap for three reasons. First, the consequences of misaligned AI are severe: an agent that optimises for transaction processing speed at the expense of fraud detection quality creates direct financial loss. Second, the regulatory environment demands explainability and auditability of decision-making, which requires formalised intent specifications, not just well-crafted prompts. Third, the industry is moving rapidly toward autonomous agent deployment in critical workflows, from credit decisioning to trade execution, where the compound effect of small misalignments can be catastrophic.

The Klarna lesson, translated to banking: an agentic AI deployed for customer complaint resolution that is measured on resolution speed might systematically offer refunds and concessions to close cases quickly, eroding margins and training customers to complain strategically. An intent-engineered system would have explicit decision boundaries: escalate high-value complaints, offer concessions only within defined parameters, and optimise for long-term customer lifetime value rather than short-term resolution metrics.

5. Active Inference: The Bridge to True Adaptive Intelligence

The limitations exposed by the intent gap, and by the broader constraints of prompt, context, and intent engineering, all point to a deeper architectural challenge. Current AI systems, regardless of how well they are prompted, contextualised, or intent-aligned, operate on a fundamental computational model designed in the 1940s. They are reactive: they receive input, process it, and produce output. They do not autonomously seek to reduce uncertainty about the world. They do not adapt their internal models in real time. They do not learn from the unexpected.

The active inference framework, developed in computational neuroscience by Karl Friston and elaborated by researchers including Pezzulo, Rigoli, and Friston (2017), offers a fundamentally different model. Active inference describes the brain as a prediction machine that continuously generates hypotheses about what it expects to perceive, then acts to minimise the difference between prediction and reality. When the prediction is wrong, the system updates its model. When the prediction is right, the system gains confidence.

5.1 The Four Principles of Active Inference

Predictive Modelling

At its core, an active inference system maintains an internal model of the world and generates continuous predictions about incoming data. This is similar to reinforcement learning, as demonstrated by DeepMind's AlphaGo, but extends beyond game environments to open-ended, real-world data streams. In financial services, this means a system that does not just react to market movements but maintains a continuously updated predictive model of market dynamics.

Perception as Hypothesis Testing

Rather than passively receiving data, the system treats every new data point as evidence for or against its current hypotheses. This principle is already applied in technologies such as Kalman Filters, widely used in GPS navigation and real-time sensor fusion. In financial services, this means treating each transaction not as an isolated event but as evidence that either confirms or challenges the system's current model of normal behaviour.

Continuous Learning

Through repeated cycles of prediction, observation, and error minimisation, the system continuously refines its model without the batch retraining cycles that current deep learning requires. This is the key difference from conventional AI: the system never stops learning, and it never becomes stale.

Action to Minimise Prediction Error

This is the genuine breakthrough. In active inference, the system does not merely react to the world; it acts to make the world more predictable and to reduce its own uncertainty. It adjusts and adapts proactively, rather than waiting for instruction. This is precisely the kind of intelligence that financial services needs for real-time risk management, adaptive fraud detection, and dynamic portfolio optimisation.

The Link to Neuromorphic Computing

Active inference describes how intelligence should work. Neuromorphic computing provides the hardware that can actually implement it. The SpiNNaker project at the University of Manchester (Furber et al., 2014) demonstrated that neuromorphic platforms, with their massively parallel, brain-inspired architectures, can natively implement the predictive and self-correcting dynamics of active inference. Spiking neural networks process information as temporal events, exactly the format required for continuous prediction and error minimisation. This is the architectural bridge between the intent gap we face today and the adaptive, self-correcting financial systems of the future. We explore this in full technical depth in our companion paper.

6. Financial Services Use Cases: Matching AI to the Problem

With the reasoning taxonomy and AI evolution framework established, we can now examine specific financial services challenges through this lens. The goal here is not to present detailed technical solutions, which are covered in our companion paper, but to clarify what different categories of problem require from an intelligence standpoint, and to highlight where current AI approaches may be better suited to some tasks than others.

6.1 Retail Banking

Customer Service and Engagement

Current LLM-based chatbots handle tier-one queries effectively: account balance enquiries, transaction disputes, and FAQ responses. However, as the Klarna case demonstrates, autonomous AI in customer service requires intent engineering to ensure it optimises for customer lifetime value rather than resolution speed. The next evolution, active inference-based systems, would continuously learn from customer interactions to predict needs before they arise.

KYC, Onboarding, and Customer Segmentation

Know Your Customer processes currently rely on a combination of rule-based checks (heuristics) and machine learning models for document verification and risk scoring. Agentic AI is beginning to automate end-to-end onboarding workflows. However, because customer risk profiles change over time, continuous reassessment is becoming increasingly important. This is an area where batch-trained models may be limited, and where active inference systems running on neuromorphic hardware could offer advantages.

6.2 Rare Event Detection: Fraud, Insider Trading, and AML

This is arguably the most important use case for advanced AI in financial services, and it illustrates why current approaches are approaching their limits.

Why Current AI Falls Short

Fraudulent transactions and insider trading are rare events, typically representing less than 0.1% of all activity, creating severely imbalanced datasets. Labelling suspected fraud cases requires expensive specialist input from compliance officers and investigators. Current machine learning models, trained on historical data, struggle to detect novel fraud techniques that differ from past patterns. Rule-based systems generate overwhelming false positive rates, with some institutions reporting that over 95% of AML alerts are false alarms, consuming thousands of analyst hours in manual review.

The fundamental problem is one of reasoning type. Fraud detection with rules engines uses deductive reasoning: it can only catch what has been explicitly anticipated. Fraud detection with machine learning uses inductive reasoning: it can generalise from past fraud patterns but fails when adversaries innovate beyond the training distribution. What fraud detection actually requires is a combination of abductive reasoning (inferring the best explanation from incomplete evidence in context) and analogical reasoning (recognising that a new fraud scheme shares structural dynamics with past schemes despite looking entirely different on the surface).

What the Problem Demands

Effective next-generation fraud detection needs four capabilities that current architectures struggle to provide simultaneously: event-driven processing that responds in milliseconds rather than batch cycles, continuous learning that adapts to new attack vectors without full retraining, graph-level analysis that examines network relationships rather than isolated transactions, and energy efficiency that allows sophisticated analytics to run at the edge rather than requiring a round-trip to a cloud data centre. These requirements point toward a fundamentally different computing paradigm, one that we explore in depth in our companion paper.

The Trader's Gut Feeling

Consider what happens every day in an investment bank's trading floor. A seasoned trader monitors market signals with extraordinary precision, running calculations, reading order flows, and analysing news. But beyond the analytics, something else is at work: experienced traders develop what they describe as a "gut feeling" about market movements, sensing something in the market that is not registered in any database. This is analogical reasoning: the brain recognising structural patterns from past experience and applying them to a novel situation. It is the one form of reasoning that current AI cannot replicate, and it is the form that neuromorphic computing, with its brain-inspired spike-timing dynamics, is designed to emulate.

6.3 Commercial and Investment Banking

In commercial and investment banking, the challenges extend beyond fraud detection to encompass real-time risk propagation, market volatility analysis, and regulatory compliance. AI systems must process unstructured data, including earnings calls, regulatory filings, geopolitical news, and social media sentiment, and synthesise it into actionable trading signals or risk assessments in sub-millisecond timeframes.

Current LLM-based systems can summarise and extract information from these sources effectively. Agentic AI can automate research workflows and draft analysis. But the reasoning required here is fundamentally analogical: understanding how a policy change in one jurisdiction might cascade through counterparty networks to affect portfolio risk in another requires recognising structural patterns across domains, exactly the type of reasoning that current inductive and abductive systems cannot reliably perform.

The additional constraint is temporal. Markets are non-linear dynamic systems where the sequence and timing of events matter as much as the events themselves. A rate announcement followed by a currency movement followed by a commodity shift tells a fundamentally different story depending on the timing gaps between them. Current AI processes data as static snapshots; what trading and risk management actually need is temporal inference, intelligence that encodes information in the timing of events, not just their occurrence. This is precisely the capability that spiking neural networks, running on neuromorphic hardware, are designed to provide.

6.4 Insurance: Networked Risk in a Connected World

Insurance underwriting and claims assessment face a challenge that mirrors fraud detection: the need to move from static, backward-looking actuarial models to dynamic, network-aware risk assessment. Traditional models assign risk scores based on historical demographic and behavioural data. But modern risk is inherently networked: a single catastrophic event can cascade through interconnected systems of policyholders, reinsurers, supply chains, and service providers in ways that static models cannot anticipate.

What insurance needs, and what current AI cannot deliver at scale, is the ability to propagate risk through a network in real time. When one node in an insurance network experiences a claim event, the system should instantly reassess the risk of all connected nodes based on the strength and nature of their connections. This is graph-level, temporal, analogical reasoning: recognising that the pattern of contagion in this event resembles the pattern in a previous event, even if the surface details are entirely different. Our companion paper presents specific performance data on how neuromorphic graph-analytics address this challenge.

Use Case	Current AI Approach	Reasoning Gap	What Is Actually Needed
Customer service	LLMs / Agentic AI	Intent misalignment (Klarna effect)	Active inference for proactive, goal-aligned engagement
KYC / Onboarding	ML + Agentic workflows	Static risk profiles from batch training	Continuous reassessment via event-driven learning
Fraud detection	ML classifiers + rules	Cannot detect novel attack patterns	Event-driven graph analytics with continuous adaptation
AML / Sanctions	Rules + ML scoring	Overwhelming false positive volumes	Temporal graph analysis that distinguishes signal from noise
Trading / Execution	Algorithmic / quant models	Static snapshots of dynamic markets	Temporal inference that encodes timing of events
Insurance risk	Static actuarial models	No network-level risk propagation	Real-time graph-based risk contagion modelling

7. The Three Walls: Why Incremental Improvement Will Not Be Enough

Throughout this paper, we have identified limitations at three distinct levels. Taken individually, each is a significant challenge. Taken together, they reveal that the current AI paradigm is approaching a structural ceiling that cannot be overcome through incremental improvement within the existing architectural framework.

Wall 1: The Reasoning Wall

Current AI systems, from rule engines to Large Language Models to agentic frameworks, can replicate three of the four types of human reasoning: deductive (rules), inductive (pattern recognition), and abductive (contextual problem-solving). But they cannot reliably perform analogical reasoning: the perception of deep structural similarity across domains that enables experienced humans to navigate novel situations. This is not a matter of scale; adding more parameters, more data, or more compute to a transformer architecture does not produce analogical reasoning, because the architecture was not designed for it. The reasoning wall is a design constraint, not a resource constraint.

Wall 2: The Intent Wall

As the Klarna case study demonstrates, even AI systems that perform their designated tasks with exceptional capability can fail catastrophically at the organisational level if they are not aligned with genuine business intent. The progression from prompt engineering to context engineering to intent engineering represents increasingly sophisticated attempts to solve this problem within the current paradigm. But intent engineering ultimately requires AI systems that can understand trade-offs, adapt to changing priorities, and self-correct when their actions produce unintended consequences. These are capabilities that demand continuous learning and active inference, capabilities that current batch-trained, static architectures cannot natively support.

Wall 3: The Infrastructure Wall

Every AI system in production today runs on the von Neumann architecture, which physically separates memory from processing. For the matrix multiplications that power neural networks, over 90% of energy is consumed moving data between memory and processor, not computing on it. The human brain, by contrast, processes the equivalent of exaflops of computation on 20 watts by computing in-memory with sparse, event-driven processing. As AI workloads grow exponentially, the energy, latency, and sustainability constraints of the von Neumann model are creating a hard ceiling on what current infrastructure can deliver. AI-driven data centre consumption is projected to reach approximately 3% of global electricity by 2030, according to IEA estimates, with Goldman Sachs forecasting a 165% increase in data centre power demand over the same period. This is not a problem that better software can solve; it requires fundamentally different hardware.

The Convergence

These three walls are not independent problems. They are manifestations of the same underlying constraint: current AI architectures were designed for calculation, not cognition. Analogical reasoning requires temporal, relational processing that von Neumann hardware cannot efficiently support. Intent alignment requires continuous learning that batch-trained models cannot provide. Sustainable scaling requires in-memory, event-driven computation that the separation of processor and memory makes impossible. The solution to all three walls is the same: a computing paradigm that mimics the biological brain. That paradigm is neuromorphic computing, and it is the subject of our companion paper.

8. What Comes Next: The Questions This Paper Opens

This paper has established a framework for understanding the current AI landscape: the types of intelligence, the types of reasoning they replicate, the generations of technology that implement them, and the critical gap between an AI system's task capability and its alignment with organisational intent. We have shown that each generation of AI addresses a progressively deeper form of reasoning, from deductive rule-following to inductive pattern recognition to abductive contextual problem-solving.

But we have also identified the frontier that current AI cannot cross. Analogical reasoning, the ability to perceive deep structural similarity between situations, remains beyond the reach of any architecture built on the von Neumann model. Active inference, the brain's mechanism for continuous prediction and self-correction, cannot be efficiently implemented on hardware that separates memory from processing and operates on rigid clock cycles.

This raises a series of questions that demand answers:

Open Questions for Paper 2

If current AI architectures cannot support analogical reasoning or active inference, what hardware paradigm can? How do spiking neural networks and memristive devices physically replicate the brain's temporal coding and in-memory computation? What does a neuromorphic financial infrastructure actually look like, and what performance can it deliver? How should financial institutions prepare today for a computing paradigm that is already demonstrating prototype-level results? What is the realistic timeline for production deployment, and what are the genuine barriers?

These questions are the subject of our companion paper, The Frontier of Cognitive Computation: Neuromorphic AI and the Future of Financial Ecosystems, in which we provide the full technical and strategic analysis of neuromorphic computing for financial services, including a three-horizon roadmap for deployment and an assessment of technology readiness using the NASA TRL framework.

References

1. Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417-424.
2. Pezzulo, G., Rigoli, F., and Friston, K. (2017). Active Inference, homeostatic regulation and adaptive behavioural control. Progress in Neurobiology, 134, pp.17-35.
3. Furber, S., Galluppi, F., Temple, S., and Plana, L. (2014). The SpiNNaker Project. Proceedings of the IEEE, 102(5), pp.652-665.
4. Vaswani, A. et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.
5. Nate Jones (2026). Klarna saved $60 million and broke its company: The missing layer is intent engineering. Nate's Newsletter / Substack.
6. Gartner (2025). Predicts 2025: AI Agents Will Transform Enterprise Operations.
7. MIT NANDA Initiative (2025). The GenAI Divide: State of AI in Business 2025.
8. Deloitte (2025). State of AI in the Enterprise, 7th Edition.
9. Neuromorphic graph-analytics engine detecting synthetic-identity fraud. World Journal of Advanced Research and Reviews (2025).
10. Neuromorphic computing: A game changer in financial planning and analysis systems. World Journal of Advanced Research and Reviews (2025).
11. Karpathy, A. (2025). On Context Engineering. Various public remarks and social media commentary.
12. Infosys (2025). Brains on Wheels: The Role of Neuromorphic Computing in Next-Gen Mobility. Infosys Emerging Technology Solutions.
13. Infosys (2025). What Agentic AI Means for Financial Services. Infosys Knowledge Institute.
14. Infosys (2025). Capitalizing on Growth: Why Financial Services Firms Need a Unified AI Strategy. HFS Research / Infosys.

About the Authors

Ricardo Garcia

Ricardo J. Garcia is a Principal at Infosys Consulting with over a decade of experience driving AI adoption and digital transformation across retail banking, corporate banking, and capital markets. He has led complex, data-driven programmes for major financial institutions including Barclays, Santander, and Computershare, spanning AI implementation, regulatory stress testing, and operational control. He holds a Master's in Financial and Management Control and a degree in Law and Economics.

Serge Plata

Serge Plata is a Senior Principal at Infosys Consulting with over 20 years' experience on developing data and AI implementations across many sectors and lately specializing on financial services. He has led technical teams and also strategy programmes for FTSE100 companies and S&P500 companies. He holds a PhD in mathematics and also has management studies at Harvard Business School.

Manglika Singh

Manglika Singh is a senior leader with 20 plus years of experience driving large scale digital transformation in Financial Services. As Associate Partner at Infosys Consulting, she leads the AI first consulting practice for EMEA Financial Services, focusing on turning AI capabilities into impactful client solutions. Her experience across Infosys, EY, and IBM iX spans technology, business strategy, and human centered design. She specializes in AI driven automation, operational efficiency, customer centric onboarding, and advancing proactive, sustainable KYC management.

ABOUT INFOSYS CONSULTING

Infosys Consulting is a next-generation consulting partner that bridges strategy and execution. With an AI-first mindset, deep industry knowledge, and the combined strengths of business and technology consulting, it helps enterprises turn bold vision into tangible outcomes, faster, smarter, and at scale. Infosys Consulting is helping some of the world's most recognizable brands transform and innovate. Our consultants are industry experts that lead complex change agendas driven by disruptive technology. With offices in 20 countries and backed by the power of the global Infosys brand, our teams help the C-suite navigate today's digital landscape to win market share and create shareholder value for lasting competitive advantage.

For more information, contact consulting@infosys.com

Infosys® | CONSULTING