THE INTELLIGENCE PROBLEM

AI agentic payments promise a continuous, 24/7 financial ecosystem operating at maximum economic efficiency, but taking the human out of the loop exposes a gaping regulatory and insurance chasm. Matthew Challis explores

For most of human history, in one way or another, we have dedicated swathes of time and energy to reducing our workload. From the invention of the wheel to the introduction of Microsoft Teams, humans have strived to take a little bit off their plates with each new technological advancement.

AI's vast and unprecedented adoption over the past five years has, in no small part, reshaped institutional finance; the technology has transitioned from the rigid, rule-based automations of yesteryear, developing into one comprising advanced neural networks and capabilities. With the increasingly prevalent shift towards a 'digitally native' financial world, one of the latest trends arrives in the form of AI agentic payments, wherein AI agents autonomously discover, negotiate, and even execute purchases on behalf of firms.

DeFi markets are defined by their programmability and 24/7 nature, which are frequently cited as the most desirable aspects of the digital assets space. Traditional cross-border B2B payments are far from frictionless. Operating delays and the constraints of TradFi working hours can result in a myriad of issues for any party at any stage of a transaction — from approval setbacks to interventions around execution.

What AI agentic payments set out to do, according to AI-powered financial operations platform Ramp, is sit somewhere between an automatic payment and a human-judged one, combining the former with the latter.

It comes as little surprise, therefore, that, with such a promising and capable technology, B2B fintech platform Galileo projects that, by 2032, the agentic payment will grow from US$7 billion to US$93 billion, and Bain & Company, a management consultancy, estimates that the US agentic commerce market could reach US$300–500 billion by 2030, representing 15–20 per cent of domestic e-commerce volume.

But without human intervention, how can institutions rest assured that their money is in safe hands? What about the decrease in security? Are the AI agents even capable of human judgment?

Most importantly, what happens when something invariably goes south, and who or what is to blame when that happens?

Judgment calls

Fundamentally, AI systems are at a crossroads with the foundational infrastructure of banks and payment networks. TradFi rails are inherently and deliberately deterministic, designed with predictable rules and fixed legal certainties, defined by binary execution methods. Conversely, AI seeks to be probabilistic, adaptive, and heuristic, with malleable decision-making and learning capabilities, intended to mimic human judgment.

Within this mishmash of operational intentions lies a great hurdle for banks: as AI is moved from pilot programmes into production, how do you integrate an adaptive intelligence into a system that has zero tolerance for errors?

In the view of Robin Hasson, global head of reconciliations at SmartStream, the solution is not to compromise on either front and to establish a clear architectural division. “Banks don’t move from pilot to production by making the rails smarter,” he argues. “They do it by making the operational layer around the rails more intelligent while keeping its outputs strictly deterministic.” Hasson believes that the instinct to push AI agents into settlement itself “is the wrong one” and advocates allowing agents to be, themselves, intelligent on rails that are kept “deliberately dumb”.

To combat the potential for errors, Hasson insists that every AI-produced proposal must clear a hard, code-enforced gate before it can carry any weight.

Tony Livesey, chief technology and product officer at AutoRek, echoes the sentiment of separated responsibilities. “The key is not to make the clearing rail probabilistic,” he warns. “The rail must remain deterministic. The AI layer should never be the final source of truth for payment, execution, entitlement, limit checking, settlement instruction, or legal finality.”

Without a plethora of mandated audits — deterministic control gates, entitlement checks, mandate validation, sanctions screening, liquidity checks, exposure limits, dual control thresholds, and immutable audit capture — Livesey believes that AI should sit not inside the clearing finality mechanism, but above it, acting as an orchestration and decision-support layer.

Itai Turbahn, vice president of embedded wallets at Fireblocks, does, too, support the notion that separation is fundamentally key if AI agents are to become a mainstay in institutional finance. He believes that agents must abide by strict rules: the agent proposes; the policy layer enforces. Only transactions that meet predefined parameters ever actually reach settlement. “The key principle is that probabilistic thinking happens upstream of the guarantee, never inside the settlement process itself,” Turbahn explains.

Autonomy or control?

In theory, AI agents sit at the intersection of a dilemma for institutional risk management teams — how much autonomy do they give the agent, and how much control can they retain for it to still serve a useful purpose?

Hasson challenges this as a false dichotomy. Instead, he believes that the “more useful idea is bounded in autonomy: the agent stays autonomous within the box, and risk owns the box”.

Conceptually, bounded autonomy affords risk officers the ability to set hard parameters without sacrificing the intelligence of the machine agent. In practice, an agent control layer enforces strict pre-trade limits, real-time exception detection, and automated circuit-breakers tied to exposure thresholds, all in a bid to limit the agent’s physical reach rather than its reasoning ability. If an agent were to propose a trade that would breach its hard-coded limits, the execution of said trade would fail before ever reaching the intended venue. The agent would then have to replan its strategy within permitted boundaries.

Hasson cites the Knight Capital disaster of 2012, in which the firm lost US$440 million in a mere 45 minutes, due to deterministic code that was executed without appropriate safeguards, such as a kill switch, in place. He warns that if AI agents exist without the necessary cautionary frameworks, it is “the same failure on a faster clock”.

The consensus among senior digital assets leadership is therefore one where policy layer controls are based on rigid, predefined rules above all else. Instead of trusting the agent wholly, autonomy is instead defined as the ability to operate effectively within clear, circumscribed boundaries. Institutions, like Fireblocks, are leveraging Multi-Party Computation (MPC) — cryptographically splitting private keys across multiple mathematical parties. In doing so, Turbahn explains, “an agent may participate in initiating an action, but it never has unilateral control of the signing process”.

The Agent Payments Protocol (AP2) — launched by Google, Coinbase, and over 60 other industry institutions — is an open-source standard designed to enable AI agents to independently transact, manage wallets, and interact with digital assets on behalf of their deployer. The protocol, which uses cryptographically signed mandates as tamper-proof evidence of a user’s given instructions, is designed to answer three critical questions surrounding authorisation, authenticity, and accountability. Its purpose is simple: to act as verifiable, undeniable proof of both user intent and authority.

For wholesale capital markets, however, AP2 permissions sit far too close to the threshold of a consumer-level checkout tool instead of something able to be fully utilised in an institutional-grade workflow or framework. Livesey argues that the industry needs to move from “the user clicked approve” to “the user granted a verifiable, revocable mandate”.

Who’s to blame?

Regulatory compliance is often unable to keep up with the pace at which the industry moves. Authoritative bodies currently mandate explicit authorisation for transaction approval. For AI agentic payments to become viable at scale, the industry must develop cryptographically signed, machine-readable policies and open protocols.

But once those protocols are configured and the keys themselves are mathematically split, who carries the liability if an autonomous agent causes a large-scale monetary loss or compliance breach?

TradFi banking rails and regulations are built on centuries-old fundamentals, with accountability rules designed for human

actors. Such a large-scale infrastructure shift leaves a gap in the regulatory frameworks and overall landscape. The lack of clarity poses itself as a major bottleneck for Tier-1 institutions. In the eyes of Livesey, without clear liability allocation, "autonomous limits will remain constrained because no board will accept open-ended ambiguity on who is responsible when an agent causes loss, breach, or market disruption".

Turbahn emphasises that, without a human in the loop, the question that remains to be seen is "who carries the risk when something goes wrong"? He believes that crypto "will be ready before the insurance and liability frameworks are" and that the gap itself is what imposes limits on human-free, unsupervised AI learning, not the technology.

Changing the regulatory landscape

Traditional compliance frameworks, like Know Your Customer (KYC) and Know Your Business (KYB), while sufficiently capable of deterring and preventing human bad actors, are ill-equipped to deal with the autonomous nature of AI agents. Without the necessary data to build a profile, agents inherently bypass a significant transaction security layer.

The verifiable Legal Entity Identifier (vLEI), promoted by the Global Legal Entity Identifier Foundation (GLEIF), has emerged as the base layer for machine recognition. It functions as a way for counterparties to computationally verify the identity and legal authority of an individual or software engineer who is acting on behalf of an institution.

However, Livesey argues that corporate identification alone is insufficient in satisfying capital market compliance, and the industry requires a layered infrastructure for Know Your Agent (KYA), the agentic extension of traditional KYC and KYB frameworks. He believes that, for KYA to be effective, it must verify a plethora of identifiers: agent designations and model; software versions and code integrity; an operator's real-world identity; permitted operational functions and the cryptographic mandate chain; real-time wallet-to-account binding; dynamic sanctions and counterparty screening; behavioural monitoring to detect model drift or anomalous transaction patterns; and real-time revocation and quarantine capabilities.

The challenge in implementing such a robust framework, Turbahn says, "is connecting existing compliance infrastructure to a new identity and authorisation model designed for autonomous systems". While the intended agentic output remains equivalent to that of a human actor, how enterprises navigate the path to reach that destination is fundamentally different. Without a plug-and-play compliance architecture, it is therefore not institutionally viable to adopt a potentially volatile transaction model.

There is another risk that lingers in the minds of compliance teams: algorithmic coupling. If, for instance, multiple Global Systemically Important Banks (G-SIBs) and buy side institutions deploy payment and trading models trained and optimised on similar or the same AI models, the market, theoretically, risks a stark and potentially dangerous loss of behavioural diversity.

Under normal market conditions, these models would likely operate with the anticipated speed and efficiency of agentic AI. Under stress, however, a hypothetical "monoculture" of AI models could trigger synchronised capital flight or self-reinforcing liquidity withdrawal, at speeds too fast for human intervention, in a manner that might not be picked up on straight away. Turbahn views this as one of the most significant dangers. "We've seen this movie before with the 2007 quant deleveraging and the algo flash crashes since," he says. "The difference now is that it happens faster, and there are fewer humans in the loop to pause and second-guess."

In response to the potential dangers and failure points associated with the technology, Livesey proposes more essential safeguards that must operate on both an institutional level and across broader market infrastructures. He advocates for model diversity, execution randomisation and throttling, pre-trade market impact controls, institutional-level concentration limits, market-level circuit breakers, and regulatory observability. Without them, the market risks another catastrophe, akin to Knight Capital.

How viable?

If the ecosystem can overcome these not-insignificant regulatory and technological hurdles, capital markets may enter a new, transformational era. A continuous, 24/7 financial ecosystem, with markets that clear almost instantaneously, at a speed close to theoretical maximum economic efficiency. Humans, for all our capabilities, are burdened with un-optimisable faults: sleep, families, social lives, office hours, and manual settlement windows. AI agents, by their very nature, do not succumb to such limitations.

But this transition does not imply "markets without humans". Rest assured, the human role will remain, although in a slightly different capacity. Instead of manual transaction execution, objective writing, and higher-level system design, the general oversight of agents will be the new, crucial role.

The future, Livesey says, belongs to "controlled autonomy". Platforms that succeed will be built on governed frameworks — such as AutoRek's ARIA governance model or SmartStream's deterministic Air reconciliation engine — ensuring that even as financial speed approaches real time, safety remains fully anchored in human hands.

And, as Turbahn points out: "Whoever writes the mandate holds the power, not whoever runs the fastest agent." ■