KLA INSIGHTS

Before Giving an AI Agent an Account, Give It a Framework for Trust and Stewardship

Jumpei Miwa | Kunitachi Liberal Arts Inc. (KLA)

Prepared for the International FinTech Forum, FinTech Taipei 2026, originally scheduled for 24 October 2026

A note for readers

The event in Taiwan has been postponed to mid-December due to an approaching typhoon. Even though there will be another opportunity to speak in December, I would like to share in the form of a speech draft what KLA wanted to communicate about AI agents this time. This is not a record of a delivered address, but a prepared draft adjusted for publication. It retains the flow of the original speech, with event-day wording and selected factual statements adjusted, and references added.

I had been looking forward to returning to this forum to discuss how we should approach AI agents in finance.

Six years ago, back in 2020, I joined this forum online as a financial regulator of the Japan Financial Services Agency when the pandemic unfolded. Despite the adversity of the pandemic preventing us from physical appearance, the digital transformation of finance accelerated worldwide. At that time, I spoke about open banking APIs, data use responsibilities in the context of AI development, risk-based payment regulations, and governance in decentralized finance.

While many of those challenges remain pressing today, those topics now sound like the classic debates of an earlier era. As technology was changing the boundaries of finance, how should we reorganize the responsibility to protect users? That was my question as a financial regulator.

Around 2020, AI conversations centered around credit scoring and fraud detection, grappling with core issues of data quality, bias, and accountability.

Today, six years later, that question has moved further. We are entering an era where AI does more than analyze and advise. It receives authority from humans and carries out actual transactions.

Suppose that tomorrow an AI agent compares products representing a customer, prepares identity verification documents, hedges foreign exchange exposure, and even makes payments, what should we regulate in an era of Agentic AI?

Is it a model? Is it a transaction? Or an authority? My answer is “Authority”.

AI policy and financial responsibility

Looking back over the past eight years of international discussion, policy of AI has evolved in three stages. The first stage focused on explainability in machine learning, data quality, and bias. The second focused on cyber risks, third-party dependencies, and model risk associated with generative AI. And now, in the third stage, we are beginning to confront the issues raised by AI agents that make their own plans, call external tools, and execute transactions.

There is a striking degree of common ground in the messages from the FSB, the IMF, the Basel Committee, and the OECD. AI has the potential to improve productivity, compliance, price discovery, and customer services. But concentration in the same models or cloud providers,

correlated decisions, cyberattacks, and models that are difficult to explain can transform risks at individual firms into systemic risks. What we need, therefore, is not to put a stringent regulation that stops AI, but to put a design that keeps the financial system operating even when AI falls into failures.

While these shared risks are widely recognized, jurisdictions differ significantly in how they address them. The EU takes a rules-based approach, adding new AI-specific legislation. The UK takes a principles-based approach, applying existing principles flexibly with a focus on outcomes. The US takes a federal approach, applying existing laws case by case. Although their methods differ, they share a firm consensus: ultimate responsibility can never be shifted to AI itself.

In the UK, the Bank of England and the FCA found in a 2024 survey that 75 percent of responding financial firms were already using AI. In Japan, the Financial Services Agency published its AI Discussion Paper in 2025 and updated it to Version 1.1 in 2026. The regulatory question is therefore shifting from whether to use AI to how to use it safely.

AI agents and on-chain infrastructure

Meanwhile, the industry is moving further ahead. Stripe is rolling out the Agentic Commerce Protocol, which enables purchases through AI platforms. Mastercard is working toward payments by verified AI agents through Agent Pay. Coinbase's x402 uses HTTP's "Payment Required" response to enable AI agents to pay API fees automatically in stablecoins.

Coinbase introduced Agentic Wallets in February 2026, with session caps, transaction limits, key isolation, and built-in transaction screening.

Against this backdrop, the significance of blockchain dramatically changes. For humans, blockchain has not been easy to use, requiring users to navigate complex wallet setups, securely store private keys and manage seed phrases, and accept the constant risk of irrevocable loss from a single operational error. But, for AI agents operating around the clock and issuing trade instructions instantly via APIs, on-chain technology could become the most native financial infrastructure and settlement layer.

As previously mentioned, in the Coinbase-led x402 protocol, an automatic '402 Payment Required' HTTP status code response triggers an AI to pay instantly for paid APIs or web content using stablecoins—starting from just a few cents per transaction, without human intervention. Fast machine-to-machine transfers of very small amounts, or micropayments, can thus be completed immediately, without credit card entry screens or complicated login procedures.

There is an exceptionally strong fit between AI handling decisions and on-chain infrastructure that provides programmable authority, settlement, and evidence of execution. In the era of AI agents and autonomous finance, building trust frameworks and on-chain infrastructure is a core strategic management agenda for financial institutions.

However, on-chain technology is not a cure-all. A ledger can prove what was executed, but not verify why the decision was correct. With flawed prompts, contaminated data, or stolen authority, an accurate ledger merely records a wrong action accurately.

In other words, while on-chain infrastructure provides a necessary foundation for automated execution, it does not provide the sufficient condition for safety and governance. Moreover, gas fees and network latency remain practical constraints for micropayments. It is precisely because of these technological and governance limits that I propose giving AI agents a framework for trust and stewardship.

Four pillars of trust and stewardship

To put this into practice, let me explain the four pillars that make this concrete.

First, Identity: Establish explicitly who the agent represents, ensuring verifiable accountability at every touchpoint.

Second, Mandate: Embed machine-readable boundaries that strictly define what it may transact, up to what limit, with whom, and for how long.

Third, Settlement: Engineer risk-proportional execution layers—balancing immediate finality with escrow, holds, and recourse.

Fourth, Evidence and Exit: Guarantee complete auditability and provide fail-safe mechanisms for humans to revoke authority, halt execution, and assume control instantly.

Industry implementations already contain elements pointing in this direction, including spending limits, agent verification, and key isolation.

This is not about disregarding the "human in the loop." Rather, it shifts our focus from reactive ex-post approvals, to a proactive, stewardship-driven governance—placing human involvement firmly at the ex-ante stage of authority design and system architecture.

For high-risk credit and investment decisions, we should strengthen human approval. For routine, low-risk payments, we should allow greater autonomy within authority defined in advance. It is important not to treat autonomy as all or nothing.

In the IMF's 2024 survey, market participants also emphasized human oversight from the perspectives of regulation, risk, responsibility, and ethics.

Testing authority and sustaining responsibility

Regulators, banks, payment providers, AI companies, and blockchain providers everywhere face the same imperative: to test AI agents with defined authority in a limited sandbox.

The question is no longer just how to build a smarter model, but what metrics truly matter—the rate of actions beyond its authority, the time needed to stop it, the ability to recover from erroneous transactions, the completeness of audit trails, and continuity during third-party failures.

Forward-looking jurisdictions around the world now have a unique opportunity to build and show a working model of this framework together, including Taiwan and Japan.

While international bodies like the IMF and FSB highlight systemic risks, such as model concentration and correlated decisions, preventing the next financial crisis requires looking beyond traditional safeguards. The key is not only to build smarter AI, but to explicitly define its boundaries in a machine-readable form. This brings us back to the core of what we must establish: trust and stewardship.

Let me close by returning to trust and stewardship. Trust is not a vague ideal; it rests directly upon structural pillars: verified identity, defined mandates, secure settlement, and verifiable evidence with exit paths. Stewardship is our unwavering responsibility to embed these safeguards into both our institutions and our technology.

In short: before giving an AI agent an account, give it a framework for trust and stewardship.

The more autonomous AI becomes, the greater our responsibility. Together with everyone here, I would like to build a financial system capable of fulfilling that responsibility.

References

1. FSB, The Financial Stability Implications of Artificial Intelligence, 14 November 2024
https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

2. IMF, Global Financial Stability Report, October 2024, Chapter 3, Advances in Artificial Intelligence: Implications for Capital Market Activities
https://www.elibrary.imf.org/display/book/9798400277573/CH003.xml

3. Basel Committee on Banking Supervision, Digitalisation of finance, 16 May 2024
https://www.bis.org/media-releases/20240516-basel-committee-publishes-report-digitalisation-finance

4. OECD, Regulatory approaches to Artificial Intelligence in finance, 2024
https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/09/regulatory-approaches-to-artificial-intelligence-in-finance_43d082c3/f1498c02-en.pdf

5. European Commission, AI Act
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

6. FCA, AI and the FCA: our approach
https://www.fca.org.uk/firms/innovation/ai-approach

7. US Treasury, Uses, Opportunities, and Risks of Artificial Intelligence in Financial Services, December 2024
https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf

8. Bank of England and FCA, Artificial intelligence in UK financial services – 2024
https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

9. Japan FSA, Publication of AI Discussion Paper, 2025; Version 1.1, 3 March 2026
https://www.fsa.go.jp/en/news/2025/20250304/aidp.html

10. Stripe, Developing an open standard for agentic commerce, 29 September 2025
https://stripe.com/blog/developing-an-open-standard-for-agentic-commerce

11. Mastercard, Mastercard unveils Agent Pay, 29 April 2025
https://newsroom.mastercard.com/news/press/2025/april/mastercard-unveils-agent-pay-pioneering-agentic-payments-technology-to-power-commerce-in-the-age-of-ai

12. Coinbase, Introducing Agentic Wallets: Give Your Agents the Power of Autonomy, 11 February 2026
https://www.coinbase.com/es-us/developer-platform/discover/launches/agentic-wallets

13. Coinbase Developer Documentation, How x402 works
https://docs.cdp.coinbase.com/x402/how-it-works

14. Coinbase Developer Documentation, CDP Facilitator
https://docs.cdp.coinbase.com/x402/seller/facilitator

日本語訳

AIエージェントに口座を与える前に、信頼とスチュワードシップの枠組みを与えよ

三輪 純平 | 株式会社国立リベラルアーツ（KLA）

2026年10月24日に開催が予定されていた「FinTech Taipei 2026」国際FinTechフォーラム向け準備原稿より

読者の皆様へ

台風の接近に伴い、台湾でのイベントは12月中旬に延期となりました。12月に再び発言の機会はありますが、今回KLAがAIエージェントに関して伝えようとしていた内容を、スピーチ原稿の形でお届けしたいと思います。本原稿は実際の講演の記録ではなく、公開用に調整された準備稿です。元のスピーチの流れを維持しつつ、イベント当日の表現や一部の事実関係の記述を調整し、参照文献を追加しています。

金融におけるAIエージェントにどのように向き合うべきかを議論するため、このフォーラムに再び戻ってくることを楽しみにしておりました。

6年前の2020年、パンデミックが拡大する中、私は日本の金融庁の規制当局者としてオンラインで本フォーラムに参加いたしました。パンデミックという試練により対面での往来が阻まれる中でも、世界中で金融のデジタル変革が加速していきました。当時私が申し上げたのは、オープンAPI、AI開発の文脈におけるデータ利用責任、リスクベースの決済規制、そして分散型金融（DeFi）のガバナンスについてでした。

それらの課題の多くは今日もなお切実ですが、今やそれらのテーマは一世代前の古典的な議論のように響きます。技術が金融の境界を変えていく中で、利用者を保護する責任をどのように再構築すべきか。それが金融当局者としての私の問いでした。

2020年当時のAIをめぐる議論の中心は信用スコアリングや不正検知であり、データの品質、バイアス、そして説明責任という根幹の課題に取り組んでいました。

6年後の今日、その問いはさらに前進しています。私たちは、AIが単に分析や助言を行うにとどまらず、人間から権限を与えられ、実際の取引を自ら執り行う時代へと足を踏み入れています。

明日にも、AIエージェントが顧客に代わって商品を比較・選定し、本人確認手続きを整え、為替リスクをヘッジし、決済まで完了させるとしたら――この「エージェント型AI（Agentic AI）」の時代において、私たちは何を対象として規制すべきなのでしょうか。

モデルでしょうか？ 取引でしょうか？ それとも「権限」でしょうか？私の答えは、明確に「権限」です。

AI政策と金融における責任

過去8年間の国際的な議論を振り返ると、AIに関する政策は三段階で進化してきました。第一段階は、機械学習の説明可能性、データ品質、そしてバイアスに焦点が当てられました。第二段階では、生成AIに伴うサイバーリスク、サードパーティへの依存、そしてモデルリスクに焦点が移りました。そして今、第三段階において、自ら計画を立て、外部ツールを呼び出し、取引を実行するAIエージェントがもたらす課題に、私たちは直面し始めています。

FSB、IMF、バーゼル銀行監督委員会、OECDから発せられるメッセージには、際立った共通点があります。AIは、生産性、コンプライアンス、価格発見機能、顧客サービスの向上をもたらす大きな可能性を秘めています。しかし、同一のモデルやクラウドプロバイダーへの集中、相互に相関した意思決定、サイバー攻撃、そして説明困難なモデルは、個別の金融機関のリスクをシステム全体のリスクへと変貌させかねません。だからこそ私たちに必要なのは、AIを停止させる厳格な「規制」ではなく、AIが障害を起こしたとしても金融システムが機能し続ける「設計」なのです。

こうした共通のリスクが広く認識されている一方で、各国・地域のアプローチには明確な違いがあります。EUはAI専用の法的枠組みを追加するルールベースのアプローチをとっています。英国は既存の原則を結果（アウトカム）重視で柔軟に適用する原則ベースのアプローチをとっています。米国は既存の法律をケースバイケースで適用する連邦レベルのアプローチをとっています。手法は異なれど、「最終的な責任をAI自体に転嫁することは決してできない」という確固たるコンセンサスを共有しています。

英国では、イングランド銀行とFCAが実施した2024年の調査において、回答した金融機関の75%がすでにAIを活用していることが明らかになりました。日本でも、金融庁が2025年に「AIディスカッションペーパー」を公表し、2026年にはVersion 1.1へと改訂を行っています。したがって規制上の問いは、「AIを使うべきか否か」から、「いかに安全に使うか」へと移行しているのです。

AIエージェントとオンチェーンインフラ

そうした中、産業界の実装はさらに先へと進んでいます。Stripeは、AIプラットフォーム経由の購買を可能にする「Agentic Commerce Protocol」を展開しています。Mastercardは、「Agent Pay」を通じて検証済みAIエージェントによる決済の実現に取り組んでいます。Coinbaseの「x402」は、HTTPの「Payment Required」レスポンスを活用し、AIエージェントがAPI利用料をステーブルコインで即座に自動決済することを可能にしています。

Coinbaseは2026年2月、セッション上限、取引限度額、鍵の隔離、取引スクリーニングを備えた「Agentic Wallets」を発表しました。

こうした背景のもと、ブロックチェーンの意義は劇的に変化します。人間にとって、ブロックチェーンは決して使いやすいものではありませんでした。複雑なウォレット設定を扱い、秘密鍵やシードフレーズを安全に保管し、たった一度の操作ミスで取り返しのつかない損失を被るリスクを常に受け入れなければならなかったからです。しかし、24時間365日稼働し、API経由で即座に取引指示を発行するAIエージェントにとって、オンチェーン技術は最もネイティブな金融インフラ、決済層となり得るのです。

先ほど触れたように、Coinbaseが主導するx402プロトコルでは、HTTPの「402 Payment Required」ステータスコードによる自動応答をトリガーとして、AIが人間の介在なしに、有料APIやWebコンテンツの料金をステーブルコインで即座に（1取引あたり数セントから）支払うことができます。これにより、クレジットカード情報の入力画面や複雑なログイン手順を経ることなく、超少額の資金移動（マイクロペイメント）を含むマシン・ツー・マシンの高速決済が即座に完了します。

意思決定を担うAIと、プログラム可能な権限、決済、そして実行の証跡を提供するオンチェーンインフラとの間には、極めて高い親和性が存在します。AIエージェントと自律型金融の時代において、信頼の枠組みとオンチェーンインフラの構築は、金融機関にとって経営戦略の核心をなすアジェンダとなります。

しかしながら、オンチェーン技術は万能薬ではありません。台帳が証明できるのは「何が実行されたか」であり、「なぜその判断が正しかったか」ではありません。不完全なプロンプト、汚染されたデータ、あるいは奪われた権限のもとでは、正確な台帳は「誤った行動」を正確に記録するに過ぎません。

言い換えれば、オンチェーンインフラは自動実行のための必要条件を満たすものの、安全性とガバナンスの十分条件を提供するものではないのです。さらに、ガス代やネットワークの遅延も、マイクロペイメントの実務上の制約として残ります。まさにこうした技術的・ガバナンス上の限界があるからこそ、私はAIエージェントに対して「信頼とスチュワードシップの枠組み」を与えることを提案するのです。

信頼とスチュワードシップの4つの柱

これを実行に移すため、具現化を図る4つの柱について説明いたします。

第一に、アイデンティティ（**Identity**）：エージェントが誰を代理しているのかを明確に確立し、あらゆる接点において検証可能な説明責任を確保すること。

第二に、マンデート（**Mandate**）：何を、いくらまで、誰と、いつまで取引できるのかを厳格に規定する機械可読な境界線を組み込むこと。

第三に、決済（**Settlement**）：即時確定性とエスクロー、保留、遡及請求権のバランスを取り、リスクに応じた実行層を設計すること。

第四に、証跡と退出（**Evidence and Exit**）：完全な監査可能性を保証し、人間が即座に権限を剥奪し、実行を停止させ、制御を引き継ぐためのフェイルセーフ機構を提供すること。

産業界での実装には、利用限度額、エージェントの検証、鍵の隔離など、すでにこの方向性を示す要素が含まれています。

これは「Human in the Loop（人間の関与）」を軽視するものではありません。むしろ、事後的な承認という受動的なアプローチから、主体的でスチュワードシップに基づいたガバナンスへと焦点を移し、権限デザインやシステム構造という事前の段階に人間の関与をしっかりと位置付けるものです。

高リスクな信用や投資の意思決定においては人間の承認を強化すべきであり、定型的な低リスクの決済においては事前に定義された権限の範囲内でより大きな自律性を認めるべきです。自律性を「全か無か」で捉えないことが重要です。

IMFの2024年調査においても、市場参加者は規制、リスク、責任、倫理の観点から人間による監督の重要性を強調しています。

権限の検証と責任の実践

世界中の規制当局、銀行、決済事業者、AI企業、そしてブロックチェーンプロバイダーが、今や同じ命題に直面しています。それは、限定されたサンドボックス内で、明確な権限を与えられたAIエージェントをテストすることです。

もはや問われているのはより賢いモデルを構築する方法だけではありません。真に重要な指標とは何か――権限を逸脱した行動の割合、停止に要する時間、誤った取引からの回復力、監査証跡の完全性、そしてサードパーティ障害時における継続性です。

世界中の先見性のある地域（台湾や日本を含む）には今、この枠組みの実用モデルを共に構築し、提示していく稀有な機会が訪れています。

IMFやFSBといった国際機関が、モデルの集中や意思決定の相関性といったシステミックリスクに警鐘を鳴らす中、次の金融危機を防ぐには従来のセーフガードを越えた視点が求められます。鍵となるのは、より賢いAIを構築することだけではなく、AIの境界線を機械可読な形式で明示的に定義することです。これこそが、私たちが確立すべき核心、すなわち「信頼とスチュワードシップ」へと再び立ち返らせてくれるのです。

最後に、信頼とスチュワードシップというテーマに戻り結びの言葉といたします。信頼とは曖昧な理想ではなく、検証されたアイデンティティ、定義されたマンデート、安全な決済、そして退出経路を伴う検証可能な証跡という構造的柱の上に直接成り立つものです。そしてスチュワードシップとは、こうしたセーフガードを制度と技術の双方に組み込んでいくという私たちの揺るぎない責任に他なりません。

要約するならば、AIエージェントに口座を与える前に、信頼とスチュワードシップの枠組みを与えよ、ということです。

AIが自律性を増すほど、私たちの責任は大きくなります。本日お集まりの皆様と共に、その責任を果たし得る金融システムを築いていきたいと考えております。