an example to show that one side don't have what it says it has to made the deal, e.g. no money. –¿ 1) use blockchain , 2) not blockchain

15.9 closing thought

just like in human histroy, we bring so many infrastructure over our histories to establish trust. Contracts, laws, Credit systems.

At this point there are no agents with an accumulating history longer than a year. The inter-agent world is still nascent in its infancy, and we want to build that infra soon before its dawn.

16 Trust at Machine Speed: Why the Inter-Agent World Needs Its Own Clearinghouse

Author: Ming-Chang Chiu

In the 1770s, London's bankers had a problem. As the volume of cheques flowing between institutions grew, so did the errors, disputes and failed settlements. Each bank sent clerks to every other bank daily, a tangle of bilateral visits that collapsed under its own complexity. The solution began not in a boardroom but in a tavern — clerks from rival banks meeting at the Five Bells on Lombard Street to exchange cheques in one place — and over the following decades crystallized into a formal institution: the clearinghouse. It solved three problems at once: it verified who was at the table (membership), it standardized how transactions were communicated (protocols) and it guaranteed that what was agreed upon was actually executed (settlement). The clearinghouse did not make bankers more honest. It made honesty *verifiable*.

I believe we are approaching an analogous moment in artificial intelligence. As AI agents begin to transact with one another on behalf of different principals, they will encounter the same fundamental problem those London bankers faced: how do independent parties, each acting in their own interest, establish enough mutual trust to do business reliably and at scale? And as with banking, the answer will not be better agents. It will be new infrastructure.

16.1 Agents are becoming peers, not just tools

Today, most AI agents operate in single-sided configurations: an agent executes a workflow, or a user interacts with an agent-powered chatbot. The dominant communication protocols — the Model Context Protocol, Google's Agent-to-Agent — are architected around client-server relationships in which a central agent orchestrates tools or delegates subtasks to subordinates. This topology implicitly solves the trust problem: a single coordinator enforces correctness.

But this topology will not hold. As deployment scales, agents representing different principals will increasingly interact directly — procurement agents negotiating with vendor agents, contracting agents coordinating across firms, and, in research, automated laboratories sourcing reagents, scheduling instrument time or coordinating multi-site experiments through agent intermediaries. When agents carry different principals' interests to the table, the relationship becomes negotiative. No privileged

coordinator guarantees that commitments are honored. No clearinghouse verifies that what was offered matches what was understood. Each pair of agents operates bilaterally, exactly as London's banks did before the clerks found their way to the Five Bells.

16.2 The language that enables oversight also enables drift

Agents communicating with other agents might be expected to converge on structured, machine-optimized protocols. But the principals behind these agents are people — researchers, managers, clinicians — who need to audit what their agents communicated, understand what was agreed to and intervene when something goes wrong. Natural language is the only medium that preserves this legibility. The pressure to maintain human interpretability will keep agents talking in our language, even when doing so introduces ambiguity. And over the long-horizon interactions that define real-world agent work — multi-step negotiations, extended contracts, iterative procurement cycles — small ambiguities compound in ways that are difficult to detect and costly to unwind.

16.3 Well-intentioned agents fail in ways security cannot catch

My research has focused on surfacing hidden failure modes that aggregate metrics obscure — subgroup fairness failures in image classification, color-contrast failures in medical imaging, behavioral inconsistencies in vision-language models. In each case, the system appeared to work. The failures were real but invisible until you knew where to look. Moving from single models to interacting agents, I see structurally similar patterns — but compounded by the fact that failures now emerge *between* systems rather than within them.

Let me make this concrete. Imagine two agents negotiating a reagent order for an automated laboratory. Agent A requests 500 units of a compound at $12 per unit. Agent B confirms "the order" — but its internal resolution maps the request to a different catalogue entry with a slightly different purity grade. Neither agent flags a discrepancy; each exchange, evaluated locally, is coherent. Over the next several exchanges, delivery timelines and payment terms are negotiated on the basis of this silently divergent understanding. Both agents produce logs showing agreement at every step. The failure surfaces only when the wrong compound arrives at the lab — or worse, when it is used in an experiment that quietly produces unreliable results.

I call this phenomenon *semantic drift*: the gradual, often undetectable divergence between what two agents believe they have agreed upon. It requires no bad actor. It emerges naturally from the interaction of ambiguous language, context-dependent interpretation and extended temporal scope. And it is not the only such failure mode. Agents can contradict their own earlier commitments without detecting the inconsistency. Surface-level metrics — response coherence, timeline adherence, confirmation signals — can all read as healthy while the underlying semantic content has quietly diverged.

These problems are amplified by a feature inherent to the substrate. Large language models are stochastic: the same agent, given the same input under slightly

different conditions, can produce different outputs. In single-agent applications, this variability is managed through temperature controls, structured outputs and human review. In *multilateral*, multi-agent interactions, non-determinism on *both sides* creates a combinatorial expansion of possible trajectories. And because agents operate at machine speed — completing in seconds what would take human negotiators days — the window for compounding failure is compressed dramatically.

Existing infrastructure does not address these failures because they are not security failures — they are trust failures. Firewalls and authentication protect against adversaries. Nothing currently protects against well-intentioned agents that simply drift.

16.4 Trust must be built, not assumed

Human societies encountered structurally similar problems and, over centuries, developed institutional responses. When merchants could not verify a trading partner's creditworthiness, credit systems emerged. When counterparties could not trust each other to settle honestly, clearinghouses were established. When verbal agreements proved unreliable at scale, contract law formalized commitments into enforceable instruments. These institutions were not built because humans are fundamentally dishonest. They were built because even honest parties need verifiable commitment mechanisms when the stakes are high and the interactions are complex.

The inter-agent world needs analogous infrastructure: third-party trust layers that sit between interacting agents and provide independent verification, monitoring and dispute resolution — a clearinghouse for the age of autonomous AI. Any such transaction follows a cycle — encounter, interaction and settlement — and the clearinghouse must serve each phase. At encounter, this means identity and credibility services that verify not just authentication but whether an agent's principal has the resources and authority it claims. During interaction, this means real-time monitoring for semantic drift and commitment contradictions — catching compounding errors before they reach settlement. At settlement, this means verifiable execution and recourse mechanisms that both parties can trust precisely because they are independent of either.

For the research community specifically, the implications extend beyond commerce. Multi-agent systems are already entering scientific workflows: orchestrating literature review, data analysis and experimental execution across specialized sub-agents. As these systems grow more autonomous and begin to coordinate across institutional boundaries — sharing data, negotiating computational resources, integrating results from distributed experiments — the same trust failures that plague commercial agent interactions will threaten scientific reproducibility. An agent-mediated analysis pipeline that silently drifts from its stated methodology is a reproducibility crisis waiting to happen, except now the provenance trail looks clean because every agent logged its steps faithfully. Trust infrastructure that monitors inter-agent consistency is not just a commercial need; it is a scientific one.

Building this infrastructure poses research challenges that no single discipline currently owns. The first is measurement: current natural language processing can assess whether a single document is coherent, but we lack metrics for quantifying how far

apart two agents' internal understandings have drifted over the course of a multi-turn exchange. Detecting that drift has occurred is insufficient; what is needed is something closer to semantic telemetry — the ability to measure, in real time, the growing distance between what each agent believes has been agreed upon, and to distinguish productive negotiation from silent divergence. The second challenge is verification under stochasticity: traditional formal verification assumes deterministic systems, but large language models are inherently probabilistic. How do we verify that a commitment has been upheld when the agent's behavior is drawn from a distribution rather than computed from fixed rules? New methods are needed to define and enforce behavioral envelopes — bounds within which an agent's outputs must remain to honor its commitments, even as individual responses vary.

The third is the tension between oversight and privacy: a trust layer must observe the interaction to verify consistency, but principals may be unwilling to expose proprietary strategies, internal prompts or sensitive data to a third-party monitor. Designing oversight architectures that can verify inter-agent consistency without requiring full transparency is an open systems problem with no current solution.

Beyond the technical, there are questions at the intersection of computer science, economics and law that remain largely unaddressed. What constitutes a 'contract' between two AI agents, neither of which is a legal person? How should liability be allocated when compounding drift — not malice — produces a material loss? If an agent's representations are probabilistic by nature, what standard of 'agreement' applies? These are not hypothetical questions. They will become urgent the moment inter-agent transactions involve real money, real goods and real consequences for the principals who deployed them.

16.5 The window is open but narrowing

The history of technology transitions offers a consistent and cautionary pattern: infrastructure lags deployment, and the gap is closed reactively. The internet scaled for a decade before institutional frameworks for digital commerce and privacy began to catch up. Social media platforms reached billions of users before content-governance frameworks were seriously attempted. In each case, the lag imposed enormous and largely avoidable costs.

As of this writing, no AI agent has an accumulating interaction history longer than roughly a year. The inter-agent world is in its infancy. We have, perhaps for the first time in a major technology transition, a window in which trust infrastructure can be designed deliberately — informed by the principles of institutional design rather than dictated by crisis response.

When London's bankers established the clearinghouse, they did not wait for the system to collapse. They saw that bilateral trust could not scale, and they built a neutral institution before the volume of transactions made the problem unmanageable. The inter-agent world needs its own clearinghouse moment — in commerce, in science, wherever agents are beginning to represent independent parties with real stakes. The question is whether the research community will recognize this, and build accordingly, before the first large-scale failures force our hand.