

Post-Consultation Position Paper: Attribution Infrastructure for Autonomous Financial Agents

Submitted to: Secretariat to the Financial Stability Board

Email: fsb@fsb.org

Date: 24 September 2026

Reference: P100626: Sound Practices for Responsible Adoption of Artificial Intelligence (AI), June 2026

Submitted by:

Eric Chan

DTech Candidate, Purdue Polytechnic Institute

chan629@purdue.edu

Note on timing: The formal consultation period closed on 22 July 2026. This paper is a post-consultation position paper offered for consideration in the final report, expected October 2026. The author acknowledges that it falls outside the formal response window.

Abstract

When an autonomous financial agent makes a consequential decision, three questions must be answerable retroactively: which agent acted, was it operating within its authorized scope, and is there a verifiable record? The challenge compounds in multi-hop delegation chains, where a human authorizes Agent1, which in turn delegates to Agent2. In these H2A2A (Human-to-Agent-to-Agent) configurations, no existing regulatory framework specifies which principal bears accountability at each delegation hop, or what calibration requirements govern agents that receive delegated authority rather than direct human authorization. The Financial Stability Board's June 2026 consultation on Sound Practices for Responsible AI Adoption (P100626) requires all three attribution elements but specifies no technical protocol for how they are cryptographically authenticated and transmitted to supervisors. The resulting gap is not organizational. It is infrastructural.

This paper names and analyzes that gap. The core distinction it introduces is between data-layer attribution (proving a record was not modified after issuance) and authorization-chain attribution (proving an agent was authorized to act within scope at the time of a specific decision). These are different infrastructure requirements; the first is largely solved by blockchain and cryptographic hashing; the second has no established standard. Analysis of thirteen regulatory documents spanning six geographic jurisdictions

confirms that Multi-Agent Attribution (the specification of accountability for delegated authorization chains) scores 0/12 across all Category A governance frameworks (pre-adjudication inter-rater reliability $\kappa = 0.6507$). The proposed framework, AIF v0.1, addresses authorization-chain attribution through three components: W3C Decentralized Identifier and Verifiable Credential specifications for agent identity and delegation-chain attestation, and a 13-field Calibration Reporting Standard (CRS) applying privacy-preserving verification to calibration governance disclosures.

The framework is grounded in practitioner evidence from three operational contexts: autonomous content policy enforcement at Meta (2017-2019), e-invoicing attribution infrastructure at a digital asset infrastructure organization in Abu Dhabi (2026), and autonomous fraud and Sybil detection at OKX (2023-2024). An implementation path for FSB member jurisdictions is proposed, with a proportionality mechanism tied to the materiality assessment framework in Sound Practice 5.

Keywords: autonomous financial agents, agentic AI governance, accountability infrastructure, H2A2A governance, multi-agent delegation, human delegation provenance, decentralized identifiers, verifiable credentials, authorization attestation, calibration reporting standard, zero-knowledge machine learning, AI accountability, financial regulation, Financial Stability Board, deployer liability

Executive Summary

Sound Practices 2, 3, 5, and 10 of P100626 establish what autonomous financial agents require: defined accountability structures, integrated risk management, materiality assessment, and human-in-command oversight. The consultation further specifies that institutions should assign individual identifiers to AI agents, log agent activities, maintain scope inventories, and treat agents as "synthetic employees" with the associated access controls and accountability assignments.

What the twelve practices do not specify is the technical substrate that makes those requirements auditable. When an autonomous agent makes a consequential decision, three questions must be answerable by an external auditor: Which agent made this decision? Was it operating within its authorized scope at the time? What is the verifiable record? Without a protocol for how that record is structured, signed, and transmitted, "human-in-command" governance is a policy intention rather than an operational control.

This paper names that gap and proposes a three-component technical framework to address it, grounded in W3C open standards and academic research currently underway at Purdue Polytechnic Institute. The accountability framework informing the analysis derives from Bovens (2007), who defines accountability as a relationship in which an actor is obliged to explain and justify conduct to a forum with the capacity to impose consequences. Applied to autonomous agents, Bovens' three-element structure (forum, information, consequences) maps directly to the technical requirements P100626

specifies: identifiers and logs (information), supervisory review (forum), and deployer liability (consequences). What the twelve practices lack is the protocol that makes the information element independently verifiable.

Section 1: What P100626 Requires and Where the Gap Lies

1.1 The governance architecture P100626 establishes

The consultation's treatment of agentic AI is one of its most useful contributions. Sound Practice 2 requires that financial institutions "define clear roles and responsibilities and maintain an appropriate governance framework to enable responsible AI adoption." For agentic systems specifically, the consultation describes a certification model: "In more mature organisations, this has taken the form of agent 'certification', where an agent is reviewed and approved for use within defined boundaries — with its certified status and permitted scope recorded as inventory attributes" (P100626, p. 27).

Sound Practice 10 adds operational content to this governance requirement. Institutions should be "assigning and documenting individual identifiers to AI agents to facilitate monitoring and identity management" and "placing boundaries on AI agents, such as defining actions that are prohibited or that require human approval at appropriate checkpoints" (P100626, p. 47). The human-in-command oversight model, described as "high-level human oversight, including deciding the extent of autonomy, setting guardrails, and managing its overall impact" (P100626, p. 46), depends on those boundaries being technically enforceable and verifiable after the fact.

The consultation also recommends monitoring and documenting the process of AI agents, specifically capturing "the reasoning at each autonomous decision point; the APIs' data and other tools accessed by the agent; and the final outcome (including the agent's interpretation of success)" (P100626, p. 48). And the document closes the human oversight section by recommending that institutions adapt "human resources controls and processes to AI agents in a way that treats them as synthetic employees" (P100626, p. 48).

This is a coherent framework. The analogy to synthetic employees is precise: employees have identity documents, defined job scopes, and activity logs. A compliance function reviewing an employee's decision can link the action to the person, confirm that person's authority to take it, and examine the record. The consultation correctly identifies that the same infrastructure must exist for AI agents.

1.2 The gap: protocol for the decision record

The consultation describes what to build (identifiers, scope inventories, activity logs) but does not specify the technical protocol for how those elements are structured, cryptographically authenticated, or transmitted to an auditor or supervisor.

This matters for a concrete reason. An auditor investigating a disputed automated credit decision or a flagged fraud enforcement action needs to establish: the agent's identity, the authorization in force at the time the decision was made, and whether the decision was within that authorization. Those three questions map directly to P100626's requirements. But a spreadsheet-based inventory of agent scope, combined with application logs in a proprietary format, does not constitute a verifiable decision record in any standardized sense. The inventory may have been accurate at certification time; it may not reflect operational scope six months later. The logs may exist; they may not be independently verifiable.

A regulator might reasonably ask: is the activity log P100626 recommends not already sufficient? The consultation's recommended log captures "the reasoning at each autonomous decision point; the APIs' data and other tools accessed by the agent; and the final outcome" (P100626, p. 48). That is a good start. But an institution-held log in a proprietary format has a fundamental limitation: it is only as reliable as the institution's infrastructure and incentives to preserve it accurately. Regulators already recognize this distinction in other contexts, requiring cryptographic receipts in payment settlement, timestamped trade records in market surveillance, and signed audit trails in custody. The attribution infrastructure gap is that no equivalent standard has been defined for autonomous agent decisions.

The consultation responses published on 6 August 2026 address organizational governance, risk management integration, model lifecycle controls, and human oversight requirements. Among the approximately 20 public responses in the August 6, 2026 publication (including submissions from major banking industry associations, globally active financial institutions, payments networks, and civil society organizations), no submission proposed a protocol-level specification for cryptographic authentication of agent identity, scope authorization, and decision records as a unified standard applicable across FSB member jurisdictions. The author acknowledges that not all responses may have been reviewed in full, and invites the FSB Secretariat to correct this characterization if any submission addresses the gap. The gap is not organizational. It is infrastructural.

Systematic review of thirteen regulatory documents across six geographic jurisdictions — the United States, United Kingdom, European Union, Singapore, Hong Kong, and the UAE — confirms the gap is not specific to any single framework. Coding those documents against Bovens' (2007) five-component accountability framework produced a consistent finding: Multi-Agent Attribution, defined as the specification of who bears accountability across delegated authorization chains, receives no affirmative coverage in

any Category A governance document reviewed (pre-adjudication inter-rater reliability $\kappa = 0.6507$, $n = 13$). Every jurisdiction reviewed requires accountability assignment for AI agents; none specifies what the attribution record must contain when authorization is delegated across agent hops.

The gap is most acute in multi-hop configurations. A compliance function certifies Agent1 for credit pre-screening; Agent1 delegates a data-retrieval sub-task to Agent2; Agent2 calls Agent3 for real-time market data. When Agent3's output influences an adverse credit decision, no current regulatory framework specifies which principal bears accountability for Agent3's calibration status, whether its scope was validly delegated through the chain, or what authorization record must exist at each hop. These Human-to-Agent-to-Agent (H2A2A) architectures are not edge cases — they are the operational structure of production multi-agent financial systems. Independent analysis confirms that governance specification is a structurally absent layer across all major agent interoperability protocols, a gap that extensions to existing identity or communication standards cannot close (Kang & Diponegoro, 2026). Section 3 proposes an attribution infrastructure that addresses both the single-agent and multi-hop delegation cases.

1.3 Sound Practice 5 and the runtime verification problem

Sound Practice 5 asks institutions to "implement an effective and systematic approach to assess the materiality and risk of AI use cases at the inception stage and thereafter" (P100626, p. 31). The "and thereafter" clause is doing significant work. A materiality assessment at inception is only as good as the institution's ability to verify, after deployment, that the agent is operating within the assessed scope.

Without a technically verifiable authorization record, the "and thereafter" requirement has no teeth. You can assess materiality at design time. What you cannot do, without an attribution infrastructure, is confirm that the deployed agent at runtime is the same agent assessed at certification, operating within the same authorized scope.

Section 2: Practitioner Evidence

Three operational contexts produced the same gap independently. All three are drawn from the author's direct professional experience; this is practitioner evidence, not a systematic cross-institutional study. The value of this framing is that it traces the gap to the specific governance decisions (what threshold, which error type, what review cadence) rather than to regulatory failure in the abstract. The convergence across three different industries, regulatory environments, and organizational structures suggests the issue is structural, not incidental to any one system. The dissertation research underway at Purdue will subject this claim to systematic cross-jurisdictional regulatory analysis.

2.1 Proxy accountability at scale (Meta, 2017-2019)

The 80% relative reduction in ad policy violations I delivered across the Greater China market between 2017 and 2019 was real, measurable, and operationally meaningful. It was also a proxy. Between 2017 and 2019, I held accountability for automated content policy enforcement at Meta, where enforcement operated at hundreds of millions of daily ad impressions. Individual decision tracing was operationally impossible at that volume. The accountability mechanism was statistical: a sample of live ad impressions estimated the platform-wide policy violation rate, and that rate was the operational measure of enforcement health.

Two governance decisions governed that system with no external requirement to document them. First: what violation rate constituted adequate enforcement. Second: which error type to prioritize. The operational sequence (enforce aggressively first, accepting higher false-positive rates early, build measurement infrastructure for false positives, then relax thresholds) was a deliberate governance choice. No external standard required disclosure of that priority ordering, the trigger criteria for moving from aggressive enforcement to threshold relaxation, or the review cadence.

The accountability structure worked at the organizational level: there was a named team and a named metric. At the individual decision level, accountability was a proxy. The distinction between organizational accountability and individual-decision accountability is precisely what the "synthetic employees" framing in P100626 creates the need to resolve.

2.2 Data-layer attribution versus authorization-chain attribution (digital asset infrastructure, Abu Dhabi, 2026)

In 2026, I was responsible for an e-invoicing attribution system built for a digital asset infrastructure organization in Abu Dhabi, under the UAE Federal Tax Authority e-invoicing program (voluntary adoption from July 2026; mandatory rollout commencing January 2027). The system used distributed ledger technology to create tamper-evident, timestamped records: originating agent signature, cryptographic hash of invoice content, independently verifiable by external parties. Auditors could confirm that a record had not been modified after issuance.

My initial framing of the accountability problem here was wrong in a useful way. I went into that work expecting the central gap to be audit tooling (richer log formats, better activity streams). What I found, building the system, was more fundamental. The blockchain record did exactly what it was supposed to do. But when the question shifted from "was this record modified?" to "was this agent authorized to issue this invoice, under what conditions, and by whom?" the infrastructure had nothing to say. We had data-layer accountability. We had no authorization-chain accountability.

The experience clarified a distinction that P100626 does not yet name. Data-layer attribution and authorization-chain attribution are different things. The blockchain record satisfied an information integrity obligation. It did not establish that the agent which created the invoice was authorized to do so, within what scope, by whom, and under what conditions. Those questions require a separate infrastructure layer. Sound Practice 2's "clear roles and responsibilities" maps to authorization-chain attribution. The activity log P100626 describes in the SP 10 section maps, at best, to data-layer attribution.

Both are necessary. Neither substitutes for the other.

2.3 Autonomous enforcement without a verifiable authorization record (OKX, 2023-2024)

Between July 2023 and March 2024, I served as principal product manager for growth products at OKX, including requirements ownership for an automated Sybil and fraud detection system protecting incentive campaigns against multi-wallet abuse. I specified the detection signals: IP address clustering, wallet address tracing, device fingerprinting, and transaction pattern analysis. Engineering built the classifier to those specifications. The system made enforcement decisions autonomously: flagging an account as Sybil blocked that account from campaign rewards, a real financial consequence with no prior human review.

When the system produced false positives, the accountability assignment was genuinely unclear. I defined the signals and thresholds. Engineering built the classifier. Neither party had an unambiguous claim on the false-positive outcome. More practically: when a flagged user asked why they had been blocked, there was no record that linked the specific enforcement decision to the specific combination of signals and threshold values that triggered it. The appeals process was underdeveloped, in part because the records needed to support a proper appeal did not exist in any retrievable, interpretable form.

This is precisely the "contestability" mechanism P100626 identifies under Sound Practice 10: "mechanisms for human appeal and redress of AI outputs" (P100626, p. 46).

Contestability requires a decision record. The decision record requires a standard.

The three cases map directly to the framework components proposed in Section 3. The Meta case (Section 2.1) motivates the CRS calibration governance fields: the absence of any external standard requiring disclosure of error-type prioritization, threshold criteria, and review cadence is exactly the gap the CRS addresses. The UAE e-invoicing case (Section 2.2) motivates the authorization-chain component of the W3C VC attestation design: data-layer integrity records existed and worked; authorization scope records did not exist. The OKX case (Section 2.3) motivates the contestability provisions embedded in the CRS: the absence of a record linking enforcement decisions to specific threshold values made structured appeals operationally impossible.

Section 3: A Technical Framework for Decision Record Attribution

The three cases share a common structural feature: an autonomous agent made consequential decisions, a human was nominally accountable, and the technical infrastructure for connecting the specific decision to the accountable human was absent or ambiguous. The following framework uses existing open standards to address that infrastructure gap.

3.1 Agent identity: W3C Decentralized Identifiers

P100626 recommends "assigning and documenting individual identifiers to AI agents to facilitate monitoring and identity management." The consultation does not specify what format those identifiers should take, which leaves open the risk that different institutions across FSB member jurisdictions develop incompatible proprietary identifier schemes.

The World Wide Web Consortium's Decentralized Identifier (DID) specification (W3C DID v1.0, 2022) provides a technically mature, open standard for assigning persistent, cryptographically anchored identifiers to entities including software agents. A DID persists across the agent's operational lifecycle, is verifiable independently of any single institution's infrastructure, and creates a stable anchor point to which authorization attestations and decision records can be linked. Adopting an existing open standard for agent identifiers would avoid interoperability fragmentation as agentic AI spreads across FSB member jurisdictions.

Concretely, a DID for a credit-scoring agent at an EU-regulated bank might take the form:

```
DID: did:web:bank.example.com:agents:credit-scoring-v3-prod
Controller: did:web:bank.example.com:compliance
Created: 2026-01-15T09:00:00Z
Verification method: bank.example.com/agents/credit-v3/keys/1
```

This identifier is institution-resolvable, version-specific (credit-scoring-v3 vs v2 are distinct agents), and anchored to the compliance function as controller. An auditor querying a disputed credit decision from March 2026 can resolve the DID to confirm which agent version was deployed, who controlled it, and whether that version was active at the time of the decision.

I have deployed this pattern in production contexts (not for credit scoring, but for invoice-issuing agents in the Abu Dhabi e-invoicing infrastructure). The DID model works for data-layer attribution. What we did not build was the authorization-chain layer on top of it. That is the gap Section 3.2 addresses. Prior work has applied W3C DID and VC specifications to agent-to-agent trust establishment in multi-agent systems (Rodriguez Garzon et al., 2026). The present paper applies the same specifications to a distinct problem: auditor-facing authorization attestation, where a deploying institution must demonstrate retroactively that a specific agent was operating within its authorized

scope at the time of a consequential decision. These are different use cases requiring different design choices.

3.2 Authorization attestation: W3C Verifiable Credentials

P100626 describes agent "certification" with "certified status and permitted scope recorded as inventory attributes." An inventory spreadsheet serves internal governance. It does not serve external verification.

The W3C Verifiable Credentials specification (W3C VC v1.1, 2022) provides a standard format for signed, cryptographically verifiable claims about an entity's properties, including authorized scope, issuing authority, and validity period. An AI agent assigned a DID and holding a Verifiable Credential issued by the deploying institution would carry, in verifiable form: the agent's authorized scope at the time of any given decision, the identity of the authorizing institution, and the validity period of that authorization.

A VC authorization attestation for the credit-scoring agent above would include, at minimum:

```
Subject DID: did:web:bank.example.com:agents:credit-scoring-v3-prod
Issuer DID: did:web:bank.example.com:compliance
Authorized scope: retail-credit-scoring
  Jurisdiction: EU
  Max exposure per decision: EUR 250,000
  Prohibited: SME lending, mortgage origination
Valid period: 2026-01-15 through 2026-07-15
Re-attestation trigger: model update or scope change
Signature: [issuer's verification key]
```

Under this model, every consequential decision is linkable to the authorization in force at the time. An auditor receiving a query about a disputed March 2026 credit decision can verify, without contacting the institution's internal systems, whether the agent held a valid VC, what scope it authorized, and whether the decision fell within that scope. This is the technical implementation of what P100626's SP 2 certification model and SP 10 human-in-command framework both require.

3.3 Calibration disclosure: ZK-ML and a proposed reporting standard

Sound Practices 3 and 5 together imply an ongoing obligation to monitor and document how AI systems perform against calibrated thresholds. P100626 notes that activity logs should capture "the reasoning at each autonomous decision point." My three practitioner cases demonstrate that the governance gap is not in the activity log itself. The gap is in the calibration metadata: what threshold triggered the decision, which error type was prioritized, what the review criteria are for adjusting those thresholds over time.

Institutions are understandably reluctant to disclose ML calibration parameters, because doing so can create gaming risk (adversarial agents adapt to published thresholds) or reveal commercially sensitive information. Zero-knowledge machine learning

approaches, an active area of academic research, allow a deploying institution to prove properties of a model's calibration (that the false-positive rate is below a stated threshold, that the training data meets a stated quality standard) without revealing the model architecture, training data, or specific threshold values. Recent work demonstrates that zero-knowledge attestation of specific calibration metrics is computationally feasible under financial regulation constraints (Uddin et al., 2026), attesting demographic-parity gap and expected calibration error under SR 11-7 for credit underwriting. The framework proposed here extends that approach from individual metric attestation to a structured governance property standard covering the full set of calibration governance disclosures a deploying institution must be able to demonstrate under FSB Sound Practices.

The Calibration Reporting Standard (CRS) proposed in this paper specifies 13 mandatory attestation fields:

#	Field	Description	Proposed initial threshold
1	Agent DID	W3C DID of the attested agent	Required
2	Reporting period	Start/end dates of coverage	Required
3	System scope	Authorized use case and jurisdiction	Required
4	Detection rate	True positive rate for the primary enforcement objective	Required (no threshold)
5	False-positive rate	Proportion of decisions incorrectly actioned	Disclose; mandatory review >5%
6	Correct rejection rate	True negative rate	$\geq 80\%$
7	Expected calibration error (ECE)	Calibration gap between predicted and realized outcomes	< 0.05 = well-calibrated; > 0.10 = mandatory recalibration review
8	Human review rate	Proportion of decisions escalated to human review	$< 0.1\%$ triggers Calibration Assurance Board assessment
9	Systematic failure disclosures	Any identified systematic bias or distributional shift	Required
10	Re-attestation events	Model updates, threshold changes, or scope changes since last report	Required
11	Attester DID	DID of the attesting party (internal or external)	Required
12	Attestation date	Date attestation was issued	Required
13	Deployment jurisdiction	FSB member jurisdiction(s) of	Required

		deployment	
--	--	------------	--

The ECE threshold (0.05 / 0.10) uses temperature scaling as the calibration benchmark, consistent with Guo et al. (2017), who establish that post-calibration ECE of 2–4% is routinely achievable for properly calibrated models. The CRS compliance gate of $ECE < 0.05$ sits at the upper bound of that benchmark range, making it ambitious but demonstrably achievable. ZK-SR117 (Uddin et al., 2026) validates ECE as a workable regulatory compliance metric for credit underwriting under SR 11-7, confirming the metric's fitness for financial calibration attestation. The mandatory recalibration threshold of $ECE > 0.10$ corresponds to the uncalibrated performance range for financial classifiers, making it a natural trigger for supervisory intervention rather than an aspirational target. The human review rate threshold (0.1%) is proposed as a starting point for regulatory dialogue. Both will be calibrated against supervisory experience across FSB member jurisdictions as the CRS research matures. These fields are structurally distinct from SR 11-7 model documentation requirements, which govern static model validation rather than the ongoing governance of autonomous agent decision processes.

My dissertation research at Purdue Polytechnic Institute is developing a Calibration Reporting Standard (CRS) that specifies which calibration governance properties a deploying institution must be able to demonstrate using privacy-preserving mechanisms. The standard addresses the governance decisions identified in the practitioner evidence above: what error threshold is accepted, which error type is prioritized, what triggers threshold review, and what the review cadence is. The CRS is under active development and will be submitted for academic review. I raise it here because SP 3 and SP 5 together create a disclosure obligation for calibration metadata, and the FSB may wish to specify what that disclosure requires in technically verifiable terms.

3.4 A note on implementation maturity

DID and VC specifications are W3C standards but their adoption in regulated financial services is early. I do not claim this framework is operationally mature at scale. What I claim is that the infrastructure gap is real, that it will not be resolved by governance frameworks alone, and that the technical components to address it already exist as open standards. The FSB is well-positioned to define the adoption expectation that converts "already available" into "expected by supervisors."

A proportionality note is appropriate. Smaller institutions deploying narrow, low-materiality AI use cases should not face the same attestation burden as G-SIBs running multi-agent trading systems at scale. The materiality and risk framework in Sound Practice 5 already provides the right proportionality lever: the stringency of the decision record standard could scale with the materiality assessment outcome. High-autonomy, high-materiality agentic systems warrant cryptographically verifiable attribution. Low-

autonomy, low-stakes systems might require only documented scope inventories. The architecture of the framework accommodates this without requiring a one-size regime.

Section 4: Implementation Path for FSB Member Jurisdictions

Phase 1: Deployer liability as the adoption catalyst

Several FSB member jurisdictions are developing deployer liability frameworks for AI systems. Linking deployer liability to the verifiability of a decision record creates a strong adoption incentive without mandating a specific technical standard. An institution that can produce a DID-linked, VC-attested decision record demonstrating that its agent acted within authorized scope has a defensible accountability trail. One that cannot faces greater exposure when a decision is disputed. This approach works with any compliant technical standard; the FSB need not mandate W3C DID/VC specifically. A concrete first step: FSB member jurisdictions developing AI deployer liability frameworks could insert a safe-harbor provision specifying that a cryptographically authenticated attribution record demonstrating an agent acted within its authorized scope constitutes *prima facie* evidence of compliant deployment. This provides an immediate adoption incentive and technology-neutral language.

For EU-regulated deployers operating high-risk AI systems, this incentive is already operative. EU AI Act Article 26 (applicable from August 2, 2026 per Article 113(2) of Regulation (EU) 2024/1689) imposes deployer obligations including: Article 26(5), which requires deployers to monitor operation on the basis of the instructions for use and inform providers about serious incidents; and Article 26(6), which requires deployers to keep logs automatically generated by the high-risk AI system for a minimum of six months. Annex III, Point 5(b) explicitly includes AI systems used to evaluate the creditworthiness of natural persons or establish credit scores within the high-risk category. A DID/VC attribution record is designed to support compliance with both obligations: the authorization attestation provides evidence for monitoring under Article 26(5); the timestamped decision record constitutes a log for the retention requirement under Article 26(6). The broader Phase 1 argument, that verifiable attribution records reduce deployer liability exposure, applies wherever FSB member jurisdictions are developing AI liability frameworks, of which the EU AI Act is the most operationally advanced example.

Phase 2: Attribution record as a supervisory expectation

As part of AI governance supervisory reviews, FSB member regulators could ask to see, for a sample of consequential autonomous decisions, an attribution record: which agent, what authorization, what scope. Institutions with DID/VC infrastructure can respond directly. Institutions without it will have an observable gap. The supervisory expectation

creates demand without a prescriptive mandate; market norms can then develop around technically compatible implementations. A concrete first step: as part of existing AI governance self-assessment exercises (such as MAS's annual AI governance self-assessment or equivalent FSB member jurisdiction processes), regulators could add a structured query requesting that institutions provide an attribution record for a sample of three consequential autonomous decisions made in the past reporting period. The query itself, without a mandate, generates institutional awareness of the gap.

No existing commercial governance platform combines DID-anchored agent identity persistent across model retraining events, VC-signed authorization records verifiable without recourse to the institution's own systems, and a structured CRS governance property standard. The DIF Know Your Agent specification (KYA-OS v1.0.0, July 2026) addresses agent identity as a technical standard; the present framework extends the full attestation stack to include authorization attestation and calibration disclosure as a unified reporting standard for regulatory purposes. MAS's Safeguards for Agentic Finance at Runtime (SAFR, July 3, 2026), developed in partnership with leading financial institutions under MAS's BuildFin.ai initiative, sets out a voluntary governance framework for AI agents in Singapore's financial sector covering authorization, runtime assessment, and record retention. SAFR and the AIF attribution infrastructure are complementary: SAFR governs what an agent does in operation; AIF provides the cryptographically verifiable evidentiary record that makes SAFR's runtime governance auditable retroactively by an external examiner.

Phase 3: Calibration reporting as part of AI model documentation

Sound Practice 5 requires materiality assessment "at the inception stage and thereafter." Sound Practice 9 addresses performance management on an ongoing basis. Both, together, create a logical home for calibration reporting requirements: the materiality assessment at inception documents the threshold choices and error-type priorities; the ongoing performance management tracks whether those choices remain appropriate. FSB member jurisdictions could incorporate calibration reporting into AI model documentation standards, consistent with how existing model risk management guidance frames model performance obligations. The joint Federal Reserve/OCC/FDIC supervisory letter SR 26-2 (April 17, 2026), which revised model risk management guidance for banking organizations, explicitly defers agentic AI: Footnote 3 states that "Generative AI and agentic AI models are novel and rapidly evolving. As such, they are not within the scope of this guidance." SR 26-2 is supervisory guidance rather than statute, and this deferral anticipates future coverage. A concrete first step for Phase 3: the FSB could recommend that member jurisdictions incorporate a calibration disclosure annex into AI model documentation guidance, structured around the CRS field definitions, as a natural extension of existing model risk management frameworks. This requires no new legislation; it slots into existing supervisory expectation infrastructure.

Conclusion

The "synthetic employees" framing in P100626 names the problem precisely. Synthetic employees need identity documents, defined job scopes, and verifiable activity records. The consultation recommends building all three. What it does not specify is the protocol.

This paper has argued that the gap is not peripheral to the Sound Practices framework; it is the infrastructure on which the framework's practical operation depends. Human-in-command governance, contestability mechanisms, and deployer liability all converge on the same technical requirement: a verifiable decision record that an external examiner can read without relying on the institution's own assertions.

W3C DID and VC standards provide the identity and authorization substrate. Privacy-preserving disclosure mechanisms address the calibration transparency tension. The CRS framework under development at Purdue aims to specify what calibration disclosure requires in verifiable terms.

This paper releases AIF v0.1 as a public working specification, available via this SSRN preprint, comprising three components: the Know Your Agent (KYA) identity layer implementing W3C DID v1.0, the Authorization Attestation layer implementing W3C VC v1.1, and the Calibration Reporting Standard (CRS v0.1) specifying 13 mandatory attestation fields. FSB member jurisdictions, technical standards working groups, and deploying institutions are invited to engage with the specification as a candidate technical standard for Sound Practice 10's identifier and scope requirements. Proportionate application would calibrate attestation requirements to SP5 materiality tiers: high-materiality deployments (systemically important functions, credit decisioning at scale) require full DID/VC attestation with external attester; medium-materiality deployments require DID/VC with internal attester; low-materiality deployments require logging only.

I am grateful for the opportunity to offer this as input to the final report and would welcome the chance to discuss the research in further detail.

Eric Chan

DTech Candidate, Purdue Polytechnic Institute

chan629@purdue.edu

24 September 2026

References

- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4), 447-468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- Financial Stability Board. (2026, June 10). *Sound practices for responsible adoption of artificial intelligence (AI): Consultation report* (P100626). <https://www.fsb.org/uploads/P100626.pdf>
- Financial Stability Board. (2026, August 6). *Public responses to consultation on sound practices for responsible adoption of artificial intelligence (AI)*. <https://www.fsb.org/2026/08/public-responses-to-consultation-on-sound-practices-for-responsible-adoption-of-artificial-intelligence-ai/>
- Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, & Federal Deposit Insurance Corporation. (2026, April 17). *SR 26-2: Revised guidance on model risk management*. <https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm>
- World Wide Web Consortium. (2022). *Decentralized identifiers (DIDs) v1.0: Core architecture, data model, and representations*. <https://www.w3.org/TR/did-core/>
- World Wide Web Consortium. (2022). *Verifiable credentials data model v1.1*. <https://www.w3.org/TR/vc-data-model/>
- Rodriguez Garzon, S., Vaziry, A., Kuzu, E. M., Gehrmann, D. E., Varkan, B., Gaballa, A., & Küpper, A. (2026). *AI agents with decentralized identifiers and verifiable credentials*. In Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026). arXiv:2511.02841. <https://arxiv.org/abs/2511.02841>
- Uddin, M. N., Orpita, R. T., Bhuiyan, E. R., & Anik, A. (2026). *ZK-SR117: A chunked zero-knowledge attestation design for aggregated fair-lending metrics, with a control mapping toward full SR 11-7 coverage* [Preprint]. arXiv:2608.02664. <https://arxiv.org/abs/2608.02664>
- Decentralized Identity Foundation. (2026, July). *Know Your Agent (KYA) specification v1.0.0*. <https://github.com/decentralized-identity/kya-os-mcp>
- Monetary Authority of Singapore. (2026, July 3). *Safeguards for Agentic Finance at Runtime (SAFR): Industry information paper*. <https://www.mas.gov.sg/news/media-releases/2026/mas-partners-industry-to-develop-safeguards-for-ai-agents-in-finance>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 70, 1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>

Kang, R., & Diponegoro, Y. (2026). *Governance gaps in agent interoperability protocols: What MCP, A2A, and ACP cannot express* [Preprint]. arXiv:2606.31498.
<https://arxiv.org/abs/2606.31498>