READING GROUP SYLLABUS

Cryptography for Agentic Commerce

Ideas and techniques for making agentic commerce trustworthy with cryptography

Christian Pfrang | 22 October 2026

Christian Pfrang Markets, Policy and Technology

Syllabus v1.0

Abstract

Agentic commerce inserts software agents between the buyer and the seller in a business transaction. Both sides delegate their agents to take certain actions and make decisions under certain constraints ("Don't spend more than USD 50!").

Two design constraints become very real: Don't lose money and don't break the law. Consider the case of a procurement agent interacting on behalf of a buyer with a supplier agent acting on behalf of a merchant.

We will explore papers and writings from cryptography, security engineering and the emergent literature in agentic commerce, including agent commerce protocols. We won't be primarily concerned with legal frameworks, but we will aim to develop mental models of the types of legal requirements on agentic commerce systems, and how technical tools can help to satisfy them.

When we reason about the assurances the principals need to make to each other or third parties (e.g. an auditor, insurer or regulator), and how their agents can deliver them we have to take into account trust assumptions, required evidence, and what information might leak to this party or any observer.

Payments mechanisms pose their own difficulties in agentic commerce. Stable coins settle quickly but make it difficult to reverse transactions, whereas traditional card rails have more settlement friction but offer sophisticated dispute mechanisms.

We want to highlight trade-offs and constraints that impact how we chose cryptographic mechanisms. Legibility versus privacy, Sybil resistance versus inclusion, atomicity versus recourse, longevity of evidence versus the move to post-quantum cryptography. The goal is to develop a better understanding of the cryptographic mechanisms necessary to integrate AI agents into commercial decision making, to gain and intuition about attack surfaces and failure modes.

NOTE ON THE USE OF AI

This document was prepared by the author with the assistance of an AI system (Claude by Anthropic).

Based on the author's thematic outline of the various dimensions of trust in agentic commerce, AI was used to structure these dimensions into eight sessions.

The abstract and the introductory texts for each session were written by the author. AI was used to check for conceptual and factual errors, and as a discussion partner for less familiar concepts.

The bulk of the suggested reading papers were found using AI-based search. Beginning with an opinionated core of recent papers on agentic commerce and a list of relevant projects, the AI was asked to produce a small number of suggested readings per thematic session, ranging from classical to recent, in order to bolster relevant intuition and include relevant techniques.

All content was reviewed and edited by the author, who takes full responsibility for it.

Bibliographic data was compiled with AI assistance and verified against primary sources (arXiv, publisher records, Crossref); links were machine-checked. The paper summaries were suggested by AI and reviewed and edited by the author.

CONTACT

christianpfrang.com LinkedIn @cpfrang @humanandtheloop Substack @humanandtheloop

Contents

1 Session 1: Principals, agents and mental models 3
2 Session 2: Delegation and attenuation 6
3 Session 3: Identity, discovery and reputation 8
4 Session 4: Payment authorization and settlement 10
5 Session 5: Fair exchange, receipts and disputes 12
6 Session 6: Selective disclosure and compliance predicates 14
7 Session 7: Verifiable governance 17
8 Session 8: Proof of being human 19
References 20

1 Session 1: Principals, agents and mental models

In this first session we introduce a set of mental models across security engineering, law, security economics and cryptography that help us frame the subsequent topics

The confused deputy model, describes the situation where an agent acts under the authority of the principal (“send the money") but uses the instructions of an attacker (“to my bank account"). This maps to the vivid structural failure mode, namely a **prompt injection** attack, where an agent acts differently than what the principal that delegated it had intended it to.

“Capabilities" make such confused deputy attacks harder, by handing an agent merely the authorization to carry out a sufficiently spelled out action. In short, the agent does not specify the bank account, but merely acts under its authority to “pay the merchant."

The language/formal threshold. Enforcing policies about resource spend or tool usage always requires a transition from an instruction, issued in natural language, to a formal structure, such as as policy evaluation tool. It doesn’t matter if the instruction is a prompt formulated by humans, or by an LLM itself. This threshold is where authorization and liability attach. The intent mandate put forward by AP2, for example, lives on the side of natural language. Pointers: Saltzer–Schroeder (complete mediation).

Identity is layered. Agent identity is most fully expressed in a hierarchy of layers from the agent itself, to the delegating person, that person’s role, a company’s internal structure up to the delegating company as legal person. Each link requires its own evidence and revocation mechanisms; see session 3 for details.

Trust economics. Adverse selection means that in a situation where buyers suspect fraud, sellers must accept a discount. The obverse situation arises from Sybil resistance, where trust is engineered by making principals pay up when they breach it. Of course this excludes those that can’t afford to pay or pledge the price of entry. Pointers: Akerlof; Friedman–Resnick.

Confidentiality versus legitimate access. Once commercial activity passes into regulated context, certain relevant information needs to be disclosed. Yet at the same time individuals want privacy safeguards, while enterprises want to protect confidential information. Cryptographic techniques can help address these challenges, but their application is not foolproof. In practice information can leak through metadata and other channels, even with flawless implementation. We discuss these issues in detail in 6.

What regulators demand. We outline patterns and long-standing regulatory demands from payments and anti-money laundering practices that constrain what protocols have a chance of passing regulatory muster. These include: strong customer authentication and the ability to stop and revoke payments (cf. Rivest). AML regulations typically demand customer due diligence, sanctions screening, record-keeping and implementing the "travel rule."

Dispute resolution versus atomicity. Card rails are backed by strong institutional recourse; stablecoins are irreversible by design with no recourse. This divergence is a key factor for practical design, because dispute resolution and loss attribution are likely to have a decisive influence on adoption. (cf. Anderson.)

Post-quantum. What ages well (hash-based constructions, lattice signatures) and what does not (curves, pairings) depends on the primitive it rests on. Shor’s algorithm, one of the flagship quantum algorithms, helps solve the factorization problem (used in RSA) and the discrete logarithm problem

(used in pairings and ECDSA) in polynomial time (Shor, 1997; Proos and Zalka, 2003). Migration is easier in more centralized environments, and harder under decentralized decision making. (cf. Bernstein–Lange.)

There is a three-step cascade of post-quantum vulnerability depending on what the cryptography is used to establish:

- **Encryption:** Encrypted information is susceptible to harvest-now, decrypt-later attacks, so anything that needs to be secret for long periods of time must be transitioned to PQ-safe encryption schemes early on.
- **Evidence:** We use signatures to prove to auditors and other third parties the nature of the information at the time a decision was made (e.g. "granted bank account because older than 18"). This means that for quantum-vulnerable signature schemes such a signature can be forged after Q-day, and therefore none of the accumulated evidence can be depended on anymore. The solution is to switch to a PQ-resistant signature scheme in time before the arrival of a quantum computer, and to anchor the existing records in a transparency log.
- **Credentials:** Switching to a PQ-safe signature or credential scheme with sufficient buffer before Q-day is enough. Credentials are typically only short-lived, so a smaller transition period is tolerable.

Trusted execution. TEEs are becoming an increasingly important building block for "safe" code execution and attestation. Yet in making use of their speed, relative convenience and increasing availability, one moves trust to the vendor, or cloud provider, and incurs security tradeoffs that include certain physical attack vectors. We discuss aspects of this in Session 7.

Verifiability versus accountability. Cryptography can be used to establish proof that something went wrong, but liability and accountability are assigned by law and, potentially, legal procedure. (cf. Anderson.)

Readings

- **Authentication in Distributed Systems: Theory and Practice** — Lampson, B. et al., 1992 — Origin of the "speaks-for" vocabulary — principals, compound principals, delegated authority; everything in Sessions 2 and 7 is downstream. — doi.org/10.1145/138873.138874
- **The Confused Deputy (or why capabilities might have been invented)** — Hardy, N., 1988 — Three pages naming the exact failure mode a prompt-injected agent exhibits; the opener for the seminar thesis. — doi.org/10.1145/54289.871709
- **Capability Myths Demolished** — Miller, M. S. et al., 2003 — Cleanest intuition for "designation = authorization"; the Irrevocability myth is the mandate-revocation question. — http://zesty.ca/capmyths/
- **The Protection of Information in Computer Systems** — Saltzer, J. H. and M. D. Schroeder, 1975 — Read only the design principles (Section I.A.3): complete mediation, least privilege, open design. The reference monitor between an LLM's proposal and the tools it may call is this 1975 idea — the mental model for how a mandate becomes a policy check becomes a tool call. — doi.org/10.1109/PROC.1975.9939
- **The Market for "Lemons": Quality Uncertainty and the Market Mechanism** — Akerlof, G. A., 1970 — Adverse selection in one worked example: under quality uncertainty bad goods drive out good ones unless guarantees, signalling or reputation intervene. First rung of the trust-economics mental model. — doi.org/10.2307/1879431

- **The Social Cost of Cheap Pseudonyms** — Friedman, E. J. and P. Resnick, 2001 — The economics behind the Sybil problem: when identities are free, newcomers must be distrusted ("pay their dues") or post collateral, and nothing cheaper works. The mental model for Sybil cost versus inclusion (Sessions 3 and 8). — doi.org/10.1111/j.1430-9134.2001.00173.x
- **Why Cryptosystems Fail** — Anderson, R. J., 1993 — Why bank card systems failed in practice: rarely the cryptography, usually the procedures around it and liability rules that put the loss on the customer. The mental model for cards versus stablecoins and for who carries the loss. — doi.org/10.1145/168588.168615
- **Post-quantum cryptography** — Bernstein, D. J. and T. Lange, 2017 — Seven-page review of what a quantum computer breaks (RSA, discrete logarithms, pairings) and what survives (hash-, lattice- and code-based schemes). The post-quantum mental model: which parts of the stack age well and which must be replaced. The linked copy is the authors' manuscript. — doi.org/10.1038/nature23461
- **Can We Eliminate Certificate Revocation Lists?** — Rivest, R. L., 1998 — Five pages on why revocation is the hard part of any credential system: let the acceptor set recency requirements and make the signer supply fresh evidence. The technical side of the regulators' stop-and-revoke demand; the mechanisms follow in Session 2. — doi.org/10.1007/BFb0055482

Artifacts

- **Agent Payments Protocol (AP2): Specification** — Google and contributors, 2025 — Google's mandate architecture: intent, cart and payment mandates as verifiable digital credentials. — https://ap2-protocol.org/specification/
- **Universal Commerce Protocol (UCP): Specification** — Google, Shopify and the UCP co-developers, 2026 — The commerce layer around AP2: catalog, cart, checkout, identity linking and order as capabilities an agent negotiates with a merchant over REST, MCP or A2A, with payment handled by AP2 mandates. — https://ucp.dev/
- **International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation: The FATF Recommendations** — Financial Action Task Force, 2026 — The AML set in one document: customer due diligence (R.10), record-keeping (R.11), virtual assets (R.15) and the Travel Rule (R.16, revised June 2025, in force by 2030 — card purchases stay exempt, stablecoin transfers do not). — https://www.fatf-gafi.org/content/dam/fatf-gafi/recommendations/fatf-recommendations-2012.pdf
- **Commission Delegated Regulation (EU) 2018/389 of 27 November 2017 supplementing Directive (EU) 2015/2366 with regard to regulatory technical standards for strong customer authentication and common and secure open standards of communication** — European Commission, 2018 — The legal definition of strong customer authentication: two independent factors and, in Article 5, dynamic linking of the authentication code to amount and payee — the regulatory twin of context binding. Articles 10–18 list the exemptions, including secure corporate payment processes. — https://eur-lex.europa.eu/eli/reg_del/2018/389/oj

2 Session 2: Delegation and attenuation

Given the problem of the “confused deputy" from the first session, it becomes clear that we need to allow for fine-grained behavioral constraints for agent behavior.

The baseline: OAuth. Today the web mostly uses bearer tokens in the form of OAuth scopes in order to express delegation and authorization for resources access. When agents come into the picture, the constraints of OAuth limit the expressiveness of enforcement mechanisms that would be desirable for agents. In this section we discuss mechanisms that fix some of those limitations. (cf. South et al.)

“Macaroons” are bearer tokens which offer the possibility to make access to a service contingent on explicit conditions. They also allow adding more restrictions whenever they are delegated further. This contrasts with standard **bearer tokens** which grant full authority to whoever holds them.

Instead of handing an agent a key, and instructing it in a prompt to only use it for specific purposes, **policy-based signatures** allow to issue a key to the agent, which can only sign the kinds of messages that conform to a policy that the principal has blessed. This means that no amount of prompt injection can “talk" the agent into doing something that the policy did not allow. Additionally it hides the specifics of the policy (e.g. max spend amounts) from the merchant.

Mandate bindings. A signed mandate can fix a message (AP2’s cart and payment mandates: this cart, this amount), a policy (policy-based signatures: any message that conforms), or a goal (AP2’s intent mandate is a signed prompt that records what was asked, not what is allowed: “buy socks”). The key step is from a goal, expressed in natural language, to a permitted tool call which we discuss in Session 1. This is a **semantic** step that cannot be crossed with a signature.

A further important consideration is, how we can monitor and enforce cumulate constraints in addition to per-transaction constraints: Our agent can spend USD 50 per transactions, but only up to USD 1000.

One way to enforce such a global constraint is to give an agent access to an **observer** which can maintain a count of the total spend so far. The issuer of a virtual credit card could serve as such an observer. Pre-funded (stablecoin) wallets are another way to enforce these types of global constraints. Another cryptographic approach is discussed in Session 8.

And lastly we explore how we can use **threshold signatures** to force an agent to come back to a cosigner every time it wants to act. This is a mechanism to enforce a human-in-the loop requirement, or require checking global constraints in yet a different way.

Revocation and kill. We look at techniques that can ensure expiry caveats, consume-once mandates (Lan et al.), cosigner withdrawal (FROST), credential status. ; the cosigner and the observer are the agent’s back-channel to the principal. (cf. Rivest and Miller’s irrevocability myth, both in Session 1.

Readings

- **Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud** — Birgisson, A. et al., 2014 — Anchor paper. Chained-HMAC bearer credentials with decentralized delegation; the caveat mechanism is “deterministic extraction” made concrete.

Formalized in authorization logic; compared to SPKI/SDSI. — https://www.ndss-symposium.org/wp-content/uploads/2017/09/04_3_1.pdf

- **Position: AI Agents Need Authenticated Delegation** — South, T. et al., 2025 — Position paper from the governance side, not cryptography: names OAuth 2.0 and OpenID Connect as the baseline, shows what agent delegation needs beyond it (agent-specific credentials, auditable chains of accountability), and treats the translation of natural-language permissions into scoping rules. The baseline every other reading in this session improves on. — https://proceedings.mlr.press/v267/south25a.html
- **SDSI — A Simple Distributed Security Infrastructure** — Rivest, R. L. and B. Lampson, 1996 — Public-key ancestor Macaroons compares itself to; linked local namespaces vs. a global hierarchy — a useful contrast to on-chain global registries (Session 3). — https://people.csail.mit.edu/rivest/pubs/RL96.ver-1.1.html
- **Policy-Based Signatures** — Bellare, M. and G. Fuchsbauer, 2014 — The "constrain the key" family: the signer can only sign messages conforming to an authority-specified policy, and the policy stays private. Replace employee with agent. — doi.org/10.1007/978-3-642-54631-0_30
- **Wallet Databases with Observers** — Chaum, D. and T. P. Pedersen, 1993 — Prescient model of the agent problem: a tamper-proof module the user cannot modify inside a workstation the user controls — security for organizations, privacy for individuals. — doi.org/10.1007/3-540-48071-4_7
- **Zero-Trust Runtime Verification for Agentic Payment Protocols: Mitigating Replay and Context-Binding Failures in AP2** — Lan, Q. et al., 2026 — The modern half: AP2 gives signature verification, binding and expiration, but retries, concurrency and orchestration break implicit assumptions; proposes context binding and consume-once mandates. — arxiv.org/abs/2602.06345
- **FROST: Flexible Round-Optimized Schnorr Threshold Signatures** — Komlo, C. and I. Goldberg, 2021 — threshold custody of the agent key — https://eprint.iacr.org/2020/852
- **Whispers of Wealth: Red-Teaming Google's Agent Payments Protocol via Prompt Injection** — Debi, T. et al., 2026 — confused-deputy demo: red-teaming AP2 via prompt injection. — arxiv.org/abs/2601.22569

Artifacts

- **x402 Delegation Extension** — Nevermined, 2026 — Card rails instead of stablecoins under x402: the facilitator holds a delegated card credential and enforces spend caps as a trusted counter. — https://nevermined.ai/docs/specs/x402-card-delegation
- **Biscuit: Authorization Tokens with Offline Attenuation** — Couprie, G. and contributors, 2021 — Bearer tokens with offline attenuation and Datalog caveats; the deployed descendant of Macaroons. — https://www.biscuitsec.org/
- **UCAN: User-Controlled Authorization Networks Specification** — UCAN Working Group, 2023 — Capability-based authorization with DIDs and cryptographically chained delegations. — https://github.com/ucan-wg/spec

3 Session 3: Identity, discovery and reputation

How does your agent pick another agent to interact with?

Agent identity, but also the identity of the merchant behind a delegated agent is key to establishing forms of trust between buyers and sellers. **Reputation scoring** is a mechanism which attaches to some form of persistent identity of merchant or agent.

Layers of identity An agent is delegated by a principal, which itself could be a person with a certain constrained authority that is part of an organisation. Each layer can have identity schemes with their own issuer, evidence trail and revocation mechanics. Agents might identify themselves by signing with private keys issued to them, but principals and organisations have personal and legal identities established through registres, KYC/B, ID and business wallets. Different types of transactions might need varying levels of evidence along this chain. transaction needs. (cf. Artifacts: vLEI, the European Business Wallet proposal.)

This question about choosing legitimate commercial counterparties reveals the tension between making required information for that determination legible, while preserving anonymity and privacy to the largest possible extent.

We study resources that evaluate existing trust scoring systems and consider approaches that make identity of agents and merchants more legible, as well as mechanisms such as **Privacy Pass** which establish legitimacy without a connection to identity at all.

We'll see that trust scores in agentic commerce are susceptible to similar failure modes as those in traditional online commerce. **"Sybil attacks"** allow an attacker to manufacture many identities and then use them to "poison" the ratings distribution. Douceur (2002) showed that without a central authority such Sybil attacks are always possible.

The price list of Sybil resistance. cf. Douceur proves that the threat of Sybil attacks can only be mitigated in two ways: a central authority or a cost per identity. Either way incurs specific costs, such as collateral or stake, proof of work, KYC/KYB, or a personhood credential (Session 8). Each of these costs excludes certain types of actors, or invalidations certain applications (micro-agents, newcomers). (cf. Friedman–Resnick and Akerlof in Session 1).

Reputation, stake, insurance. We discuss three ways to establish trust: a (statistical) track record, collateral that can be slashed, recourse against an issuer or insurer. Each of them requires input from the identity layer above. Card networks risk scoring mechanisms are the incumbents in this space. (cf. Hu & Rong's stake category; the ERC-8004 study.)

Readings

- **Can Trustless Agents Be Trusted? An Empirical Study of the ERC-8004 Decentralized AI Agent Ecosystem** — Xiong, X. et al., 2026 — Anchor paper. Identity, reputation and validation registries measured across Ethereum, BSC and Base; most registrations are placeholders, reputation rarely grounded in verifiable interactions, heavy coordinated Sybil behavior. — arxiv.org/abs/2606.26028
- **The Sybil Attack** — Douceur, J. R., 2002 — The impossibility result behind those findings: without a logically central authority, Sybil attacks are always possible. — doi.org/10.1007/3-540-45748-8_24

- **The EigenTrust Algorithm for Reputation Management in P2P Networks** — Kamvar, S. D. et al., 2003 — Canonical transitive-trust construction; read against the ERC-8004 paper to see what a registry would need to be grounded in. — doi.org/10.1145/775152.775242
- **Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System** — Resnick, P. and R. Zeckhauser, 2002 — Commerce-side empirical baseline: feedback almost always positive yet predictive of performance, with reciprocation and retaliation. The same pathologies, now on-chain. — doi.org/10.1016/S0278-0984(02)11030-3
- **Privacy Pass: Bypassing Internet Challenges Anonymously** — Davidson, A. et al., 2018 — Deployed answer to "is this a legitimate client?" without identity: OPRF-based anonymous tokens redeemable for access. Directly relevant to agent access tokens. — doi.org/10.1515/popets-2018-0026
- **Inter-Agent Trust Models: A Comparative Study of Brief, Claim, Proof, Stake, Reputation and Constraint in Agentic Web Protocol Design — A2A, AP2, ERC-8004, and Beyond** — Hu, B. and H. Rong, 2025 — Shared six-category vocabulary (brief, claim, proof, stake, reputation, constraint); evaluates A2A, AP2 and ERC-8004 on security, privacy, latency/cost and Sybil resistance. — arxiv.org/abs/2511.03434

Artifacts

- **x401: HTTP Proof Requirement Protocol** — Buchner, D. et al., 2026 — Sibling of x402 on the HTTP 401 side: a route refuses until the agent presents a credential proof (personhood, membership, accreditation) via the Digital Credentials API or OpenID4VP — who is paying, next to x402's paying. — https://x401.id
- **ERC-8004: Trustless Agents** — De Rossi, M. et al., 2025 — Identity, reputation and validation registries for agents on Ethereum; the spec behind the session's anchor paper. — https://eips.ethereum.org/EIPS/eip-8004
- **HTTP Message Signatures (RFC 9421)** — Backman, A. et al., 2024 — Signing HTTP messages: the integrity layer that agent requests ride on. — https://www.rfc-editor.org/rfc/rfc9421
- **Trusted Agent Protocol: Specifications** — Visa, 2025 — RFC 9421 as a card network ships it: the agent signs each request under a key in Visa's directory (the RFC's tag parameter carries Visa's profile, aligned with the IETF Web Bot Auth work), the merchant checks freshness, nonce and signature — recognition of the agent through a central registry, the Sybil price paid to Visa. The payment leg is Session 4's. — https://developer.visa.com/docs/trusted-agent-protocol/trusted-agent-protocol-specifications
- **verifiable LEI (vLEI) Ecosystem Governance Framework v4.0: Primary Document** — Global Legal Entity Identifier Foundation (GLEIF), 2026 — The legal-person layer in production: GLEIF root → qualified issuer → legal entity → official role or engagement-context role, every credential issued and revoked under one framework; thirteen pages, and no agent in them — that is the point. — https://www.gleif.org/organizational-identity/become-a-vlei-issuer-qvi/vlei-ecosystem-governance-framework/2026-03-25_vlei-egf-v4.0-primary-document_v1.2_final.pdf
- **Proposal for a Regulation of the European Parliament and of the Council on the establishment of European Business Wallets** — European Commission, 2025 — The legal-person and mandate layers as the Commission would build them: owner identification on the EUID, mandates (Art. 3(19)), role-to-attribute mappings that must be verifiable, auditable and revocable (Art. 6(2)(b)); recital 28 names agentic AI as a use case. — https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025PC0838

- **Trusted Identities for AI Agents: An Opportunity for Europe** — WE BUILD Consortium, 2026 — Three-page multi-stakeholder non-paper (Bosch, Google, Visa Europe, Ericsson, Bundesanzeiger, NL Chamber of Commerce) asking the EU to link AI agents to the wallet framework; the pitch, not a model — it contains no layer split. — https://www.webuildconsortium.eu/news/non-paper-trusted-identities-for-ai-agents-an-opportunity-for-europe-8396
- **Know Your Agent: Building the Compliance Layer for Agentic Finance** — McWhirter, S. and D. Montes Serralde, 2026 — KYA from an exchange's policy desk: five layers — agent identity and provenance, accountable principal, mandate and permissions, delegation and monitoring, revocation and auditability — the layered model of this session written as a compliance programme; the pitch, labelled as such. — https://www.binance.com/en/square/post/372666808420252

4 Session 4: Payment authorization and settlement

Agentic commerce becomes consequential, because it moves money and value.

Traditional commerce deals with a lot of the uncertainty, and opportunity for fraud by building in **friction**. Compliance checks, chargebacks, rejections, all happen at the level of **centralized gate keepers** with the ability to handle these procedures at scale.

This model faces two challenges when agents start transacting. Firstly, agentic payment networks must likely solve the same problems related to fraud and dispute handling that traditional commerce offloaded to banks and card issuers, in order to be attractive to users. And they will increasingly face the equivalent regulatory demands, which they largely escape today. Secondly, irrespective of the degree of centralization, the **speed and multitude of agent-to-agent decision making is likely to swamp the rails** of commerce that have been built for the internet as it exists today.

Regulatory questions. Traditional e-commerce solves for the question of whether a buyer is indeed an authorized payer. Agentic commerce has to solve the additional question of whether an agent has the authority to pay (cf. Waller). There is a regulatory question about to what extent an agent-initiated payment is still customer-initiated, and how to prove this fact. Fraud-liability allocation is a similarly open question: Who pays, when the agent buys the wrong thing. The design of dispute mechanisms becomes important to the extent that stablecoins become more widely adopted in agentic commerce. And payments regulation includes AML obligations.

Agentic tokens. Existing providers such as Visa, Mastercard and Stripe have been extending their payment token frameworks to agentic use cases. This amounts to implementing the mechanisms of Session 2 (attenuation, context binding, observer, revocation) by a trusted central party. Once recorded, the traces of these mechanisms can serve as the basis for statistical risk scoring mentioned in Session 3 (cf. Artifacts: the EMVCo tokenisation framework and agentic draft; Mastercard Agent Pay; the Agentic Commerce Protocol with Stripe's shared payment token. Visa's Trusted Agent Protocol, the recognition layer, is listed in Session 3.)

Where the loss lands. What happens if a payment went to the wrong counterparty, or an agent exceeded its mandate. Card rails offer chargeback and issuer liability versus theoretical finality and no recourse on stablecoin rails. In practice, stablecoin providers can and do freeze transactions, and are even obliged to be able to do so by the Genius act.

In addition to considering the **x402** agentic payment protocol, we will read about precursors and how they tried to structure around micropayments and delayed (netted) settlement.

Readings

- **Five Attacks on x402 Agentic Payment Protocol** — Li, Z. et al., 2026 — Anchor paper. Five concrete attacks that show weaknesses in authorization, binding, replay protection, and web-layer handling. — arxiv.org/abs/2605.11781
- **PayWord and MicroMint: Two Simple Micropayment Schemes** — Rivest, R. L. and A. Shamir, 1997 — Hash-chain micropayments minimizing public-key operations; long-term broker relationships, transient user–vendor relationships. The broker is the x402 facilitator, thirty years early. — doi.org/10.1007/3-540-62494-5_6
- **Micropayments for Decentralized Currencies** — Pass, R. and a. shelat abhi, 2015 — Probabilistic (lottery) payments on any ledger, usable without infrastructure changes; — doi.org/10.1145/2810103.2813713
- **SoK: Layer-Two Blockchain Protocols** — Gudgeon, L. et al., 2020 — The map of payment and state channels, commit-chains and refereed delegation: off-chain authenticated transactions with the chain as dispute recourse. — doi.org/10.1007/978-3-030-51280-4_12
- **Free-Riding the Agentic Web: A Systematic Security Analysis of x402 Payments** — Ling, S. et al., 2026 — Independent second x402 security analysis; contrast piece or student-presented alternative to the anchor. — arxiv.org/abs/2605.30998
- **SoK: Blockchain Agent-to-Agent Payments** — Zhang, Y. et al., 2026 — SoK of blockchain agent-to-agent payments along a four-stage lifecycle (discovery, authorization, execution, accounting); the map the two x402 papers sit in. — arxiv.org/abs/2604.03733

Artifacts

- **x402: An Open Standard for Internet-Native Payments** — Coinbase, 2025 — Coinbase's open standard reviving HTTP 402: stablecoin payments as a header exchange between client, server and facilitator. — https://www.x402.org/x402-whitepaper.pdf
- **x402 Protocol Specification** — Coinbase, 2025 — The protocol spec itself: header, facilitator API, payment schemes. — https://github.com/coinbase/x402
- **EIP-3009: Transfer With Authorization** — Kim, P. J. et al., 2020 — Transfer with authorization: the gasless signed-transfer primitive x402 settles with. — https://eips.ethereum.org/EIPS/eip-3009
- **The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments** — Poon, J. and T. Dryja, 2016 — Payment channels netting many small payments into few on-chain settlements. — https://lightning.network/lightning-network-paper.pdf
- **Machine Payments Protocol (MPP) Specifications** — Tempo Labs and Stripe, 2026 — The "Payment" HTTP authentication scheme; the main contender next to x402. — https://github.com/tempoxyz/mpp-specs
- **EMV Payment Tokenisation Specification — Technical Framework** — EMVCo, 2026 — The card rails' own cryptography: token service provider, domain-restricted tokens, per-transaction cryptograms, token lifecycle — the machinery every agentic card product builds on. Free EMVCo account needed for the document. — https://www.emvco.com/emv-technologies/payment-tokenisation/

- **Industry Perspectives on the Evolution of EMV Payment Tokenization** — Pandy, S. and M. Crowe, 2019 — The free overview of the framework: what a token service provider and a token requestor do, domain restriction, token assurance, cryptograms, and the move to card-on-file; written against v2.0, and the roles it explains are the ones the agentic card products reuse. — https://www.bostonfed.org/-/media/Documents/PaymentStrategies/evolution-of-payment-tokenization.pdf
- **EMV Agentic Payments — Framework for Specifications** — EMVCo, 2026 — EMVCo's first word on delegated purchasing authority over card rails: "Intent Services" as a shared registry of what the consumer authorised, complementing rather than replacing cryptographic assurance; KYA and agentic transaction indicators flagged for later publications; names the EMV specifications it will touch. — https://www.emvco.com/resources/emv-agentic-payments-framework-for-specifications-v1-0-draft-comment-period-ends-30-sept-2026/
- **Scaling Agentic Commerce with Trust** — Mastercard, 2025 — Mastercard Agent Pay: Agentic Tokens that "build upon proven tokenization capabilities", submitted in standard card fields with a dynamic verification code; agents registered and verified, recognised through Web Bot Auth over RFC 9421. — https://www.mastercard.com/global/en/news-and-trends/stories/2025/lady-gaga-dance-contest-abracadabra-fan-version/scaling-agentic-commerce-with-trust.html
- **Agentic Commerce Protocol (ACP): Specification** — OpenAI and Stripe, 2026 — The Agentic Commerce Protocol (OpenAI and Stripe): checkout plus a delegate-payment specification under which a PSP issues a scoped token — Stripe's shared payment token is bound to one seller with a currency, a maximum amount and an expiry, and is deactivated as consumed, expired, resolved or revoked. The attenuation and revocation of Session 2, performed by the PSP. — https://github.com/agentic-commerce-protocol/agentic-commerce-protocol
- **Payments in the Age of AI Agents** — Waller, C. J., 2026 — A Fed governor's framing: the question shifts from proving that a buyer is an authorized payer to proving that an agent has the authority to pay; agent-assisted versus agent-delegated commerce; liability and fraud models to be recalibrated. — https://www.federalreserve.gov/newsevents/speech/files/waller20260929a.pdf
- **How Agentic AI Will Reshape Payments** — Davidovic, S. and H. Tourpe, 2026 — IMF staff view of the stack as intent, authorization and settlement layers, and the demand that agents "preserve the deterministic requirements of payment infrastructure". — doi.org/10.5089/9781513533308.068

5 Session 5: Fair exchange, receipts and disputes

In traditional commerce **card issuers and banks deal with fraud, non-delivery and service quality issues by allowing chargebacks**.

Stablecoin payments are generally irreversible, even though issuers can freeze funds. So how will agents that use stablecoins handle payment disputes?

The mechanisms of recourse. Card network rules defines the conditions under which a cardholder can have a payment reversed. On (theoretically) irreversible (stablecoin) rails different mechanics are needed such as escrow, an issuer's freeze (a lawful-order channel, not a dispute channel) or nothing; (cf. Artifacts: the Visa rules and dispute guide, PSD2's liability articles, the GENIUS Act's

lawful-order clause.)

The receipt. When agents transact they will likely largely buy forms of digital services. How does one **ensure that payment and service delivery are linked** with each other (atomicity)? And what happens if a service was not rendered to sufficient standard? We could define a receipt that includes bindings to payment (such as a transaction identifier), the deliverable (hash of the good, or attested execution of the service), the mandate and the policy version the agent acted under (Sessions 2 and 7) (cf. Murdoch and Anderson's principles). Note that regulators do not deem the provider's record to be sufficient proof of authorisation (cf. PSD2 Article 72). As an example of the corresponding institutional versions, cf. Visa's compelling evidence rules. Ultimately a receipt can only verify facts, while legal rules assign the loss (verifiability versus accountability, Session 1).

We consider various exchange mechanisms and how they break.

Readings

- **Optimistic Fair Exchange of Digital Signatures** — Asokan, N. et al., 1998 — The classic: either both parties get the other's item or neither does, with the trusted third party involved only on cheating or crash. The facilitator-as-optimistic-TTP framing for x402 falls out of this. — doi.org/10.1007/BFb0054156
- **Zero-Knowledge Contingent Payments Revisited: Attacks and Payments for Services** — Campanelli, M. et al., 2017 — Fair exchange without a TTP; the key distinction that ZKCP suits goods, not services — hence zero-knowledge contingent *service* payments. Agents mostly buy services. — doi.org/10.1145/3133956.3134060
- **WI Is Not Enough: Zero-Knowledge Contingent (Service) Payments Revisited** — Fuchsbauer, G., 2019 — Breaks the previous item's fixes: the buyer can still learn about the good — or, in the service variant, whether the service was performed — without paying, because the witness-indistinguishable proofs it relies on are weaker than zero knowledge. The exact guarantee of the proof system decides whether the exchange is fair. — doi.org/10.1145/3319535.3354234
- **A402: Binding Cryptocurrency Payments to Service Execution for Agentic Commerce** — Li, Y. et al., 2026 — Modern attempt at end-to-end atomicity across execution, payment and delivery; TEE-assisted adaptor signatures so payment finalizes iff the service was correctly executed and delivered. — arxiv.org/abs/2603.01179
- **Security Protocols and Evidence: Where Many Payment Systems Fail** — Murdoch, S. J. and R. Anderson, 2014 — The institutional side of disputes: EMV does not produce adequate evidence; five principles for evidence-producing payment systems. The bridge to why one might choose card rails. — https://discovery-pp.ucl.ac.uk/id/eprint/10074689

Artifacts

- **Visa Core Rules and Visa Product and Service Rules** — Visa, 2026 — The chargeback as a protocol: Section 11 (dispute resolution) and the dispute conditions of categories 10–13 (fraud, authorization, processing errors, consumer disputes — non-delivery and "not as described" among them), with the time limits and the liability shifts; the rulebook the issuer applies when it acts as the optimistic third party. Public edition; read the dispute section only. — https://usa.visa.com/dam/VCOM/download/about-visa/visa-rules-public.pdf
- **Visa Dispute Management Guide for Merchants: Best Practices, Tools and Resources** — Visa, 2026 — The readable version: the dispute lifecycle, the four categories and their

conditions, and Compelling Evidence 3.0 — the list of what a merchant must produce to defeat a fraud dispute, that is, what evidence the card rails actually accept. The institutional answer to the receipt question. — https://usa.visa.com/dam/VCOM/global/support-legal/documents/merchants-dispute-management-guidelines.pdf

- **Directive (EU) 2015/2366 of 25 November 2015 on payment services in the internal market (PSD2)** — European Parliament and Council, 2015 — The liability rules behind the chargeback: Article 72 puts the burden of proof on the provider and says the record of the instrument's use "shall in itself not necessarily be sufficient" to prove authorisation or the payer's fraud — Murdoch and Anderson's point as law; Article 73, refund by the next business day; Article 74 caps the payer's liability at EUR 50, and at zero where no strong customer authentication was applied. To be replaced by the Payment Services Regulation (political agreement November 2025, final compromise text April 2026). — https://eur-lex.europa.eu/eli/dir/2015/2366/oj
- **Guiding and Establishing National Innovation for U.S. Stablecoins Act (GENIUS Act)** — United States Congress, 2025 — The stablecoin side: a permitted issuer must have the "technological capability to comply . . . with the terms of any lawful order" (Section 4(a)(6)(B)), a lawful order being one to seize, freeze, burn or prevent the transfer of stablecoins (Section 2(16)); nothing on consumer disputes or reversal. Freeze is a law-enforcement channel, not a recourse channel. — https://www.govinfo.gov/content/pkg/PLAW-119publ27/pdf/PLAW-119publ27.pdf

6 Session 6: Selective disclosure and compliance predicates

In session 3 we discussed that ensuring the legitimacy of commercial counterparties establishes a **need for legibility** which **contrasts with the desire for anonymity** which both counterparties may harbour.

When agents act on behalf of their principals in regulated contexts features that enable selective disclosure become requirements, instead of merely risk-management tools.

Disclosure, proof, and unlinkability All three properties are desirable, but satisfiable in practical applications only to varying degrees. We may want to (or have to) disclose a claim, prove a predicate over it, and keep two presentations of the claim unlinkable. There is also the question of "for whom" the presentation remains unlinkable (towards verifiers, towards the issuer, against both colluding). One way to achieve these goals is using **signed and salted hashes** of some claim. But this approach discloses values, and an agent can in theory be tracked across merchants, because the hash and signature remain the same ("linkable"). cf. Camenisch–Lysyanskaya (CL) and BBS satisfy all three; Frigo–shelat shows how to prove predicates over a standard signature;

Zero Knowledge Proofs are a natural tool that addresses these challenges. The difficulty lies in the fact that the data under consideration must be certified by a third party, for a zero knowledge proof about an entity to carry any value. **Anonymous credentials** use zero-knowledge technology, but require that the issuers use non-standard signing mechanisms which has limited their practical usefulness.

We investigate how this problem can be solved in the context of the European **eIDAS framework** where governments anchor credential data using standard signing mechanisms and individuals prove claims about the data (e.g. "is older than 18") with their wallets, without revealing it fully.

Compliance predicates Beside "older than 18": "may buy this product from this supplier within the mandate, while not revealing the price ceiling". AML requirements can be seen as predicates and disclosures: sanctions as non-membership (Privacy Pools), the Travel Rule as plain disclosure of originator and beneficiary, KYA as "is there an accountable principal behind this agent?"

How it breaks: three correlation channels. Identifiers (the static hash and issuer signature, a reused key); status checks (revocation and status lists, phone-home to the issuer); transaction metadata (amounts, timing, counterparties). Meiklejohn as the worked example of what metadata alone yields on a pseudonymous ledger; the cf Cryptographers' Feedback, §2, on why the EU wallet's salted hashes stay linkable and batch issuance only half-helps.

The containers. The container decides which cryptography is available: salted-hash disclosure is native to SD-JWT VC and mdoc, BBS and ZK suites only to the W3C model, whose BBS suite is still a draft. The EU wallet mandates the first two and makes the third optional (European Commission, 2026, Sections 5.4.1 and 7.4.3.5). By the commission's own admission this means that linking "cannot be fully eliminated" if verifiers and the issuer collude. Furthermore: a credential bound to the principal's phone hardware cannot be presented by an agent at all, so agents need delegated presentation or credentials issued to their own key.

Schemes that age well. How each scheme fares into a post-quantum world depends on the primitive it rests on. Shor's algorithm, one of the flagship quantum algorithms, helps solve the factorization problem (used in RSA) and the discrete logarithm problem (used in pairings and ECDSA) in polynomial time. All of the above-mentioned schemes are thus at risk in a post-quantum world: Chaum and Camenisch–Lysyanskaya through RSA, BBS and privacy pools due to their reliance on pairings, Frigo–shelat through ECDSA; in salted-hash disclosure the hash itself is not at risk, but a signature typically relies on ECDSA (or RSA) and is thus vulnerable.

This chapter only touches on the evidence and the credential steps of our post-quantum vulnerability cascade and the caveats mentioned in session **??** apply. However, for the case of unlinkable credentials based on zero-knowledge proofs such as the ones implemented with BBS/CL or Frigo–shelat, the PQ-transition is still an active research area.

For BBS and Camenisch–Lysyanskaya there exists a lattice-based signature scheme outlined in Argo et al. to replace them in a post-quantum safe way. It runs at a much larger proof length, and has not been standardised. Due to the construction of BBS and CL credentials not even a quantum computer can unmask any of their properties.

Readings

- **Security without Identification: Transaction Systems to Make Big Brother Obsolete** — Chaum, D., 1985 — The intuition paper for the pillar: pseudonyms per organization, credentials transferred between them. — doi.org/10.1145/4372.4373
- **An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation** — Camenisch, J. and A. Lysyanskaya, 2001 — First practical unlinkable multi-show credentials without issuer involvement, plus optional anonymity revocation — the unlinkability-vs-revocation dial in one paper. — doi.org/10.1007/3-540-44987-6_7
- **Revisiting BBS Signatures** — Tessaro, S. and C. Zhu, 2023 — algebraic structure that enable efficient proofs of knowledge with partial disclosure. — doi.org/10.1007/978-3-031-30589-4_24
- **Anonymous Credentials from ECDSA** — Frigo, M. and a. shelat abhi, 2024 — The most practical item in the pillar: deployable without changing issuers, devices or assumptions; ZK

over SHA-256 and ISO mdoc; selected for the EU age-verification spec. The eIDAS bridge. — https://eprint.iacr.org/2024/2010

- **Blockchain Privacy and Regulatory Compliance: Towards a Practical Equilibrium** — Buterin, V. et al., 2024 — The sanctions example as a membership proof: a ZK proof that funds do or do not originate from known (un)lawful sources via association sets. — doi.org/10.1016/j.bcra.2023.100176
- **A Fistful of Bitcoins: Characterizing Payments among Men with No Names** — Meiklejohn, S. et al., 2013 — The worked example for the metadata channel: no identities anywhere, yet two clustering heuristics and a handful of purchases from known services name the clusters and trace the large flows — what transaction metadata yields when the proofs hide everything else. — doi.org/10.1145/2504730.2504747
- **Practical Post-Quantum Signatures for Privacy** — Argo, S. et al., 2024 — The lattice counterpart of CL and BBS: a signature with efficient protocols, implemented, with presentations under 80 KB against a few hundred bytes for BBS. The post-quantum anonymous credential exists and runs, and is standardised nowhere; read for the shape and the sizes, not the proofs. — doi.org/10.1145/3658644.3670297

Artifacts

- **European Digital Identity Wallet: Architecture and Reference Framework** — European Commission, 2026 — The EU wallet architecture (v3.0.0, July 2026). Its §5.4 table lines up the three containers — mdoc and SD-JWT VC mandatory, both with salted-hash proofs, the W3C model optional — and its Discussion Topic G on zero-knowledge proofs is the Commission's own map: what salted hashes plus batch issuance give, what only zero knowledge gives, and the requirement that any ZK scheme work over already-issued mdoc and SD-JWT VC. — https://eudi.dev/
- **Selective Disclosure for JSON Web Tokens (SD-JWT)** — Fett, D. et al., 2025 — Selective disclosure by salted hashes over JWT claims — the pragmatic, linkable baseline, now RFC 9901; the credential profile on top, SD-JWT VC, is still a draft and is one of the EU wallet's two formats. — https://www.rfc-editor.org/rfc/rfc9901
- **Verifiable Credentials Data Model v2.0** — W3C, 2025 — The third container, and the one that separates the data model from the securing mechanism: embedded Data Integrity proofs (EdDSA, ECDSA, BBS suites) or enveloping JOSE/COSE. The BBS suite — the only unlinkable one — is still a Candidate Recommendation Draft (September 2026). — https://www.w3.org/TR/vc-data-model-2.0/
- **Cryptographers' Feedback on the EU Digital Identity's ARF** — Baum, C. et al., 2024 — Sixteen cryptographers' consensus to the Commission on ARF 1.4, fourteen pages: two of them on why salted-hash disclosure in mdoc and SD-JWT is linkable across presentations and why batch issuance only partly helps; BBS as the fix; a post-quantum section that separates short-lived authentication from long-lived evidence. — https://github.com/eu-digital-identity-wallet/eudi-doc-architecture-and-reference-framework/issues/200
- **Electronic Signatures and Trust Infrastructures (ESI); Selective Disclosure and Zero-Knowledge Proofs Applied to Electronic Attestation of Attributes; Part 1: Feasibility Study** — ETSI, 2025 — The standards body's map of the session, well beyond eIDAS: schemes (salted hashes, multi-message signatures, programmable ZK) → formats (SD-JWT VC and mdoc on the first, W3C Data Integrity with BBS on the second) → protocols; Annex A compares every scheme on unlinkability, predicates and quantum safety in one table, clause 8 covers the

privacy of status checks, Annex E is a primer on the W3C and SD-JWT containers. Read the executive summary and Annex A; the rest is reference. — https://www.etsi.org/deliver/etsi_tr/119400_119499/11947601/01.03.01_60/tr_11947601v010301p.pdf

7 Session 7: Verifiable governance

The most convenient way for a principal to specify the behavior of their agents is via a **policy language**. This enables them to outline and manage general policies, instead of giving each agent specific instructions.

Two problems arise: Firstly, how to ensure that the agent's **actions are indeed constrained by the policies** expressed in the language. Note that the constraints cannot be enforced on an LLM-based agent itself, but only on the (deterministic) harness through which the LLM accesses tools.

This is where **trusted execution environments** come in. They can host the policy engine, as well as issue a signed authorization if the requested action conforms to policy.

And secondly, how to **prove to the principal and to third parties that it indeed acts accordingly**, perhaps without revealing the specific constraints? This is where the TEE vendor's **root of trust** comes in. A vendor-linked hardware attestation binds a hash of the policy checker code to the authorization key. And for each specific action authorization a hash of the actual policy parameters is signed by that key.

Three claims, three mechanisms. The assurance splits into three claims with three different mechanisms: that the policy is legitimate (provenance: who blessed it, signed, versioned, logged); that it was correctly evaluated (verifiable computation: attestation or zero-knowledge proof); and that it was enforced (a TEE-held key, a policy-based signature or a threshold cosigner, Session 2 — so an action that did not pass the check cannot be signed. In Session 5 we defined a receipt which could also include the policy version and the signed authorization (cf. See and Tan's compliance wrapper)

The attacker model. Working with TEEs for attestation has advantages, but comes with trust assumptions regarding the entity that operates the TEE. We consider an example (cf. TEE.fail) where physical access plus root on the host allows to extract the vendor's attestation key. A TEE attestation proves to a third party that the vendor and the operator "say so."

ZK versus TEE. A proof of policy evaluation moves the correct-evaluation claim from hardware to mathematics at the cost of prover time and of a policy language that compiles to a circuit. But in exchange both vendor and operator leave the trusted base.

However, **what is the value to a third party that an agent action conforms to the principal's policy**, if the policy itself is not revealed?

One case in which this is useful is an **auditor demanding ex-post proof** that the principal had "reasonable" policies in place. This is in many cases the standard to which regulated entities are held. **The principal can hand a regulator their policy parameters** and the TEE-authorization including the signed hashes of the policy checker and the applicable parameters. This confirms both that the action under consideration was allowed by the policy, and that the given policy parameters were applicable at the time.

Preserving audit records The bottleneck is root of trust: vendor roots and attestation keys expire,

policy-signing keys rotate. Audit evidence must survive expiring keys and key rotations. This is in full generality an extremely difficult problem that in practice can only be solved by pragmatism: "Do the reasonable thing."

Readings

- **Proof-Carrying Authentication** — Appel, A. W. and E. W. Felten, 1999 — Origin of "submit a proof with your request": Taos, SPKI, SDSI and X.509 in one logic; no decision procedure, so clients submit proofs. The modern ZK designs are this plus zero knowledge. — doi.org/10.1145/319709.319718
- **A Proof-Carrying Authorization System** — Bauer, L. et al., 2001 — Deployed follow-up: goals, sessions, modules; parts of the policy hidden from unauthorized clients — the policy-privacy angle. — https://www.cs.princeton.edu/techreports/2001/638.pdf
- **Cedar: A New Language for Expressive, Fast, Safe, and Analyzable Authorization** — Cutler, J. W. et al., 2024 — The policy-language leg: expressive, fast, safe and analyzable authorization, with a sound and complete logical encoding proved in Lean. — arxiv.org/abs/2403.04651
- **Verifying Computations without Reexecuting Them** — Walfish, M. and A. J. Blumberg, 2015 — The intuition for "prove the policy check ran", written for a general computer-science audience. — doi.org/10.1145/2641562
- **Toward Confidential Cloud Computing** — Russinovich, M. et al., 2021 — Why the cloud wants TEEs and what they rest on: hardware root of trust, attestation, the hypervisor out of the trust boundary; by the people who ship it, so read its residual-risk passages — physical attacks on memory integrity, the provider still running the attestation service — as the vendor's own list. — doi.org/10.1145/3454122.3456125
- **An Exploratory Study of Attestation Mechanisms for Trusted Execution Environments** — Ménétrey, J. et al., 2022 — Short comparative study of attestation (SGX, TrustZone, SEV, RISC-V); enough to read a cloud attestation document critically. — arxiv.org/abs/2204.06790
- **TEE.fail: Breaking Trusted Execution Environments via DDR5 Memory Bus Interposition** — Chuang, J. et al., 2026 — The attacker model made concrete: a sub-$1000 DDR5 interposer plus root on the host extracts the vendor's attestation key and forges quotes the vendor's own verifier accepts — the operator with physical access is the adversary, and the attestation then proves what the operator says. — doi.org/10.1109/SP63933.2026.00101
- **Compliance-Aware Agentic Payments on Stablecoin Rails** — See, K. and X. W. Tan, 2026 — The "rail evaluates and attests" case: an x402-style payment passes through an on-chain policy wrapper that evaluates sanctions and source-of-funds rules at execution and records a PASS/FAIL/PENDING attestation; by MAS and IMDA staff writing in a personal capacity. Ask which of the three claims that attestation carries. — arxiv.org/abs/2605.00071

Artifacts

- **Pay with Confidence: How Solv Labs Built Verifiable, Auditable Agent Payments on Amazon Bedrock AgentCore Payments** — Duffy, P. et al., 2026 — The design under discussion: zero-knowledge proof of policy evaluation plus TEE attestation for agent payments; the evidence record, signed inside a Nitro Enclave, is anchored to a public chain before settlement. — https://aws.amazon.com/blogs/machine-learning/pay-with-confidence-how-solv-labs-built-verifiable-auditable-agent-payments-on-amazon-bedrock-agentcore-payments/
- **The Security Design of the AWS Nitro System** — Amazon Web Services, 2024 — AWS's

own operator model: "by design the Nitro System has no operator access" — the claim that moves trust to the vendor; enclave attestation is explicitly out of its scope, which is the next entry. — https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html

- **Nitro Secure Module API: The Attestation Process** — Amazon Web Services — What a Nitro attestation document contains and how it is signed. — https://github.com/aws/aws-nitro-enclaves-nsm-api/blob/main/docs/attestation_process.md
- **Certificate Transparency (RFC 6962)** — Laurie, B. et al., 2013 — Transparency logs: the append-only anchor pattern for governance evidence. Obsoleted on paper by RFC 9162 (2021, Experimental, undeployed); the logs in service since 2025 keep its Merkle tree and SCTs and change only the read API. — https://www.rfc-editor.org/rfc/rfc6962
- **Zero Trust for AI Agents: A Security Framework for Deploying Autonomous AI Agents in the Enterprise** — Anthropic, 2026 — The vendor's deployment guide: the session's mechanism as an operations checklist — tool gating outside the agent, append-only logs, attested identity at the top tier — without proofs or receipts. — https://claude.com/blog/zero-trust-for-ai-agents
- **BSI TR-02102-1: Cryptographic Mechanisms: Recommendations and Key Lengths** — Bundesamt für Sicherheit in der Informationstechnik, 2026 — The timeline for the lifecycle thread: classical signatures recommended until end-2035, classical key agreement until end-2031. — https://www.bsi.bund.de/dok/TR-02102-en

8 Session 8: Proof of being human

Amid a proliferation of non-human economic actors, how do you prove to others that you are, indeed, human? Or how does an agent prove that it was indeed delegated by a human to perform a certain action, while allowing for Sybil-resistance by ensuring that the same human isn't simultaneously behind thousands of such agents?

One credential per human. The advantage of "proof-of-personhood credentials" is that as a prevention mechanism for Sybil attacks it is relatively cheap. One sign-up can validate any number of agents. But it does exclude everyone who does not want to go through the personhood verification mechanism, be it through evidence by document, by phone, or Orb.

We will read about **policy-level** issues, when such distinctions might become important and what they entail regarding the disclosure of personal information.

Then we will consider several approaches to defining **"proof-of-personhood"**. We'll look at how **World ID** is implementing this using biometrics secured by ideas from multi-party computation.

And lastly, we will consider how to enforce global constraints without any central party managing the counter with anonymous tokens, thereby resolving an open question from session 2. This mechanism allows to ensure that a unique human cannot exceed a fixed total number of requests, without them being tracked.

World ID uses multi-party computation to check iris uniqueness without holding iris codes, and the example of a nationwide sugar-beet auction in Denmark allows us to imagine what private agent-to-agent negotiations could look like in the future.

Readings

- **Personhood Credentials: Artificial Intelligence and the Value of Privacy-Preserving Tools to Distinguish Who Is Real Online** — Adler, S. et al., 2024 — Policy-level framing: credentials to prove you are a real person without disclosing personal information, and which need not be biometric; the two requirements are one credential per person per issuer and unlinkable pseudonymity. Sets up the World discussion without presupposing its answer. §3.3 is the agent case: a credential holder's agent signals a real principal without naming one, and linking several agents to one principal makes a shared controller harder to hide. — arxiv.org/abs/2408.07892
- **Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies** — Borge, M. et al., 2017 — The short original: binds physical entities to virtual identities with accountability while preserving anonymity. — https://infoscience.epfl.ch/record/301744
- **Identity and Personhood in Digital Democracy: Evaluating Inclusion, Equality, Security, and Privacy in Pseudonym Parties and Other Proofs of Personhood** — Ford, B., 2020 — Comparative survey of proof-of-personhood approaches (biometric, social-graph, in-person); the right lens for evaluating World. — arxiv.org/abs/2011.02412
- **How to Win the Clone Wars: Efficient Periodic n-Times Anonymous Authentication** — Camenisch, J. et al., 2006 — Closes the loop to Session 2's budget puzzle: anonymous authentication at most n times per period, with cheaters identified via e-cash techniques — a cryptographic spend cap. — doi.org/10.1145/1180405.1180431
- **Secure Multiparty Computation Goes Live** — Bogetoft, P. et al., 2009 — The one MPC paper worth keeping as pure intuition: a nationwide sugar-beet double auction with no single party ever seeing the bids. A private agent-to-agent negotiation, 2008. — doi.org/10.1007/978-3-642-03549-4_20

Artifacts

- **World Whitepaper** — World Foundation, 2026 — Four-part series (updated March 2026). Part 2, *Achieving Proof of Human*: the Orb, iris codes, the anonymized multi-party uniqueness check and World ID. Part 1 proposes revocable delegation of a person's proof of human to an agent, bounded so that one human cannot stand behind many bots — the mechanism is left open. — https://whitepaper.world.org/
- **Large-Scale MPC: Scaling Private Iris Code Uniqueness Checks to Millions of Users** — Bloemen, R. et al., 2024 — How World checks iris-code uniqueness across parties without any one of them holding the codes. — https://eprint.iacr.org/2024/705

References

Adler, S., Z. Hitzig, S. Jain, et al. (2024). *Personhood Credentials: Artificial Intelligence and the Value of Privacy-Preserving Tools to Distinguish Who Is Real Online*. arXiv: 2408.07892.

Akerlof, G. A. (1970). "The Market for "Lemons": Quality Uncertainty and the Market Mechanism". In: *The Quarterly Journal of Economics* 84.3, pp. 488–500. DOI: 10.2307/1879431. https://www.sfu.ca/~allen/Ackerlof.pdf (visited on Oct. 21, 2026).

Amazon Web Services (2024). *The Security Design of the AWS Nitro System*. AWS Whitepaper, first published November 2022, updated 15 February 2024. https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html (visited on Oct. 22, 2026).

Amazon Web Services (2026). *Nitro Secure Module API: The Attestation Process*. https://github.com/aws/aws-nitro-enclaves-nsm-api/blob/main/docs/attestation_process.md (visited on Sept. 4, 2026).

Anderson, R. J. (1993). "Why Cryptosystems Fail". In: *Proceedings of the 1st ACM Conference on Computer and Communications Security (CCS '93)*, pp. 215–227. DOI: 10.1145/168588.168615. https://www.cl.cam.ac.uk/~rja14/Papers/wcf.pdf (visited on Oct. 21, 2026).

Anthropic (2026). *Zero Trust for AI Agents: A Security Framework for Deploying Autonomous AI Agents in the Enterprise*. eBook, 35 pp, May 2026. https://claude.com/blog/zero-trust-for-ai-agents (visited on Oct. 22, 2026).

Appel, A. W. and E. W. Felten (1999). "Proof-Carrying Authentication". In: *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS)*, pp. 52–62. DOI: 10.1145/319709.319718. https://www.cs.princeton.edu/~appel/papers/says.pdf (visited on Sept. 4, 2026).

Argo, S., T. Güneysu, C. Jeudy, G. Land, A. Roux-Langlois, and O. Sanders (2024). "Practical Post-Quantum Signatures for Privacy". In: *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Also: Cryptology ePrint Archive, Paper 2024/131. ACM, pp. 1523–1537. DOI: 10.1145/3658644.3670297. https://eprint.iacr.org/2024/131 (visited on Oct. 22, 2026).

Asokan, N., V. Shoup, and M. Waidner (1998). "Optimistic Fair Exchange of Digital Signatures". In: *Advances in Cryptology – EUROCRYPT '98*. Vol. 1403. Lecture Notes in Computer Science. Springer, pp. 591–606. DOI: 10.1007/BFb0054156. https://eprint.iacr.org/1997/015 (visited on Sept. 4, 2026).

Backman, A., J. Richer, and M. Sporny (2024). *HTTP Message Signatures (RFC 9421)*. https://www.rfc-editor.org/rfc/rfc9421 (visited on Sept. 4, 2026).

Bauer, L., M. A. Schneider, and E. W. Felten (2001). *A Proof-Carrying Authorization System*. Tech. rep. TR-638-01. Princeton University, Department of Computer Science. https://www.cs.princeton.edu/techreports/2001/638.pdf (visited on Sept. 4, 2026).

Baum, C., O. Blazy, J. Camenisch, J.-H. Hoepman, E. Lee, A. Lehmann, et al. (2024). *Cryptographers' Feedback on the EU Digital Identity's ARF*. Statement to the European Commission, June 2024, 14 pp. https://github.com/eu-digital-identity-wallet/eudi-doc-architecture-and-reference-framework/issues/200 (visited on Oct. 22, 2026).

Bellare, M. and G. Fuchsbauer (2014). "Policy-Based Signatures". In: *Public-Key Cryptography – PKC 2014*. Vol. 8383. Lecture Notes in Computer Science. Full version: Cryptology ePrint Archive, Paper 2013/413. Springer, pp. 520–537. DOI: 10.1007/978-3-642-54631-0_30. https://eprint.iacr.org/2013/413 (visited on Sept. 4, 2026).

Bernstein, D. J. and T. Lange (2017). "Post-quantum cryptography". In: *Nature* 549.7671, pp. 188–194. DOI: 10.1038/nature23461. https://pure.tue.nl/ws/files/118214786/20170630_tonature.pdf (visited on Oct. 21, 2026).

Birgisson, A., J. G. Politz, Ú. Erlingsson, A. Taly, M. Vrable, and M. Lentczner (2014). "Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud". In: *Network and Distributed System Security Symposium (NDSS)*. Internet Society. https://www.ndss-symposium.org/wp-content/uploads/2017/09/04_3_1.pdf (visited on Sept. 4, 2026).

Bloemen, R., B. Gillespie, D. Kales, P. Sippl, and R. Walch (2024). *Large-Scale MPC: Scaling Private Iris Code Uniqueness Checks to Millions of Users*. Cryptology ePrint Archive, Paper 2024/705. https://eprint.iacr.org/2024/705 (visited on Sept. 4, 2026).

Bogetoft, P., D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, et al. (2009). "Secure Multiparty Computation Goes Live". In: *Financial Cryptography and Data Security (FC 2009)*. Vol. 5628.

Lecture Notes in Computer Science. Springer, pp. 325–343. DOI: 10.1007/978-3-642-03549-4_20. https://eprint.iacr.org/2008/068 (visited on Sept. 4, 2026).

Borge, M., E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, and B. Ford (2017). "Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies". In: *IEEE Security & Privacy on the Blockchain (IEEE S&B)*, pp. 23–26. https://infoscience.epfl.ch/record/301744 (visited on Sept. 4, 2026).

Buchner, D., A. Lemmon, and D. Louie (2026). *x401: HTTP Proof Requirement Protocol*. Standard Draft 0.2.0, Proof. https://x401.id (visited on Oct. 21, 2026).

Bundesamt für Sicherheit in der Informationstechnik (2026). *BSI TR-02102-1: Cryptographic Mechanisms: Recommendations and Key Lengths*. Technical Guideline, version 2026-01, 23 January 2026. https://www.bsi.bund.de/dok/TR-02102-en (visited on Oct. 22, 2026).

Buterin, V., J. Illum, M. Nadler, F. Schär, and A. Soleimani (2024). "Blockchain Privacy and Regulatory Compliance: Towards a Practical Equilibrium". In: *Blockchain: Research and Applications* 5.1. DOI: 10.1016/j.bcra.2023.100176.

Camenisch, J., S. Hohenberger, M. Kohlweiss, A. Lysyanskaya, and M. Meyerovich (2006). "How to Win the Clone Wars: Efficient Periodic *n*-Times Anonymous Authentication". In: *Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS)*. Full version: Cryptology ePrint Archive, Paper 2006/454, pp. 201–210. DOI: 10.1145/1180405.1180431. https://eprint.iacr.org/2006/454 (visited on Sept. 4, 2026).

Camenisch, J. and A. Lysyanskaya (2001). "An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation". In: *Advances in Cryptology – EUROCRYPT 2001*. Vol. 2045. Lecture Notes in Computer Science. Springer, pp. 93–118. DOI: 10.1007/3-540-44987-6_7. https://eprint.iacr.org/2001/019 (visited on Sept. 4, 2026).

Campanelli, M., R. Gennaro, S. Goldfeder, and L. Nizzardo (2017). "Zero-Knowledge Contingent Payments Revisited: Attacks and Payments for Services". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Also: Cryptology ePrint Archive, Paper 2017/566, pp. 229–243. DOI: 10.1145/3133956.3134060. https://eprint.iacr.org/2017/566 (visited on Sept. 4, 2026).

Chaum, D. (1985). "Security without Identification: Transaction Systems to Make Big Brother Obsolete". In: *Communications of the ACM* 28.10, pp. 1030–1044. DOI: 10.1145/4372.4373.

Chaum, D. and T. P. Pedersen (1993). "Wallet Databases with Observers". In: *Advances in Cryptology – CRYPTO '92*. Vol. 740. Lecture Notes in Computer Science. Springer, pp. 89–105. DOI: 10.1007/3-540-48071-4_7. https://chaum.com/wp-content/uploads/2021/12/Wallet_Databases.pdf (visited on Sept. 4, 2026).

Chuang, J., A. Seto, N. Berrios, S. van Schaik, C. Garman, and D. Genkin (2026). "TEE.fail: Breaking Trusted Execution Environments via DDR5 Memory Bus Interposition". In: *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 4408–4426. DOI: 10.1109/SP63933.2026.00101. https://tee.fail/files/paper.pdf (visited on Oct. 22, 2026).

Coinbase (2025a). *x402 Protocol Specification*. https://github.com/coinbase/x402 (visited on Sept. 4, 2026).

Coinbase (2025b). *x402: An Open Standard for Internet-Native Payments*. https://www.x402.org/x402-whitepaper.pdf (visited on Sept. 4, 2026).

Couprie, G. and contributors (2021). *Biscuit: Authorization Tokens with Offline Attenuation*. https://www.biscuitsec.org/ (visited on Sept. 4, 2026).

Cutler, J. W., C. Disselkoen, A. Eline, S. He, K. Headley, M. Hicks, et al. (2024). "Cedar: A New Language for Expressive, Fast, Safe, and Analyzable Authorization". In: *Proceedings of the ACM on Programming Languages* 8.OOPSLA1. arXiv: 2403.04651.

Davidovic, S. and H. Tourpe (2026). *How Agentic AI Will Reshape Payments*. IMF Note 2026/004. International Monetary Fund. DOI: 10.5089/9781513533308.068. https://www.imf.org/-/media/files/publications/imf-notes/2026/english/insea2026004.pdf (visited on Oct. 21, 2026).

Davidson, A., I. Goldberg, N. Sullivan, G. Tankersley, and F. Valsorda (2018). "Privacy Pass: Bypassing Internet Challenges Anonymously". In: *Proceedings on Privacy Enhancing Technologies* 2018.3, pp. 164–180. DOI: 10.1515/popets-2018-0026.

De Rossi, M., D. Crapis, J. Ellis, and E. Reppel (2025). *ERC-8004: Trustless Agents*. Ethereum Improvement Proposal, Draft, created 13 August 2025. https://eips.ethereum.org/EIPS/eip-8004 (visited on Oct. 21, 2026).

Debi, T., W. Zhu, and P. Sen Gupta (2026). *Whispers of Wealth: Red-Teaming Google's Agent Payments Protocol via Prompt Injection*. arXiv: 2601.22569.

Douceur, J. R. (2002). "The Sybil Attack". In: *Peer-to-Peer Systems (IPTPS 2002)*. Vol. 2429. Lecture Notes in Computer Science. Springer, pp. 251–260. DOI: 10.1007/3-540-45748-8_24. https://www.microsoft.com/en-us/research/?p=154502 (visited on Sept. 4, 2026).

Duffy, P., R. Ansari, H. Shadab, M. S. Vangara, C. Shriyan, and W. Benno (2026). *Pay with Confidence: How Solv Labs Built Verifiable, Auditable Agent Payments on Amazon Bedrock AgentCore Payments*. https://aws.amazon.com/blogs/machine-learning/pay-with-confidence-how-solv-labs-built-verifiable-auditable-agent-payments-on-amazon-bedrock-agentcore-payments/ (visited on Oct. 22, 2026).

EMVCo (2026a). *EMV Agentic Payments — Framework for Specifications*. Version 1.0 Draft, released for public review 31 August 2026 (comment period ended 30 September 2026). EMVCo Agentic Payments Task Force. https://www.emvco.com/resources/emv-agentic-payments-framework-for-specifications-v1-0-draft-comment-period-ends-30-sept-2026/ (visited on Oct. 21, 2026).

EMVCo (2026b). *EMV Payment Tokenisation Specification — Technical Framework*. Version 2.4, 9 July 2026; public-access document, free EMVCo account required. https://www.emvco.com/emv-technologies/payment-tokenisation/ (visited on Oct. 21, 2026).

ETSI (2025). *Electronic Signatures and Trust Infrastructures (ESI); Selective Disclosure and Zero-Knowledge Proofs Applied to Electronic Attestation of Attributes; Part 1: Feasibility Study*. Technical Report ETSI TR 119 476-1 V1.3.1. August 2025, 163 pp. ETSI. https://www.etsi.org/deliver/etsi_tr/119400_119499/11947601/01.03.01_60/tr_11947601v010301p.pdf (visited on Oct. 22, 2026).

European Commission (2018). *Commission Delegated Regulation (EU) 2018/389 of 27 November 2017 supplementing Directive (EU) 2015/2366 with regard to regulatory technical standards for strong customer authentication and common and secure open standards of communication*. https://eur-lex.europa.eu/eli/reg_del/2018/389/oj (visited on Oct. 21, 2026).

European Commission (2025). *Proposal for a Regulation of the European Parliament and of the Council on the establishment of European Business Wallets*. COM(2025) 838 final, 19 November 2025. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025PC0838 (visited on Oct. 21, 2026).

European Commission (2026). *European Digital Identity Wallet: Architecture and Reference Framework*. Version 3.0.0, 21 July 2026. https://eudi.dev/ (visited on Oct. 22, 2026).

European Parliament and Council (2015). *Directive (EU) 2015/2366 of 25 November 2015 on payment services in the internal market (PSD2)*. Articles 71–74: notification, evidence, and liability for unauthorised payment transactions. https://eur-lex.europa.eu/eli/dir/2015/2366/oj (visited on Oct. 22, 2026).

Fett, D., K. Yasuda, and B. Campbell (2025). *Selective Disclosure for JSON Web Tokens (SD-JWT)*. RFC 9901, Proposed Standard, November 2025. https://www.rfc-editor.org/rfc/rfc9901 (visited on Oct. 22, 2026).

Financial Action Task Force (2026). *International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation: The FATF Recommendations*. As amended June 2026. https:

//www.fatf-gafi.org/content/dam/fatf-gafi/recommendations/fatf-recommendations-2012.pdf (visited on Oct. 21, 2026).

Ford, B. (2020). *Identity and Personhood in Digital Democracy: Evaluating Inclusion, Equality, Security, and Privacy in Pseudonym Parties and Other Proofs of Personhood*. arXiv: 2011.02412.

Friedman, E. J. and P. Resnick (2001). "The Social Cost of Cheap Pseudonyms". In: *Journal of Economics & Management Strategy* 10.2, pp. 173–199. DOI: 10.1111/j.1430-9134.2001.00173.x. https://hdl.handle.net/2027.42/71559 (visited on Oct. 21, 2026).

Frigo, M. and a. shelat abhi (2024). *Anonymous Credentials from ECDSA*. Cryptology ePrint Archive, Paper 2024/2010. https://eprint.iacr.org/2024/2010 (visited on Sept. 4, 2026).

Fuchsbauer, G. (2019). "WI Is Not Enough: Zero-Knowledge Contingent (Service) Payments Revisited". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Also: Cryptology ePrint Archive, Paper 2019/964. DOI: 10.1145/3319535.3354234. https://eprint.iacr.org/2019/964 (visited on Sept. 4, 2026).

Global Legal Entity Identifier Foundation (GLEIF) (2026). *verifiable LEI (vLEI) Ecosystem Governance Framework v4.0: Primary Document*. Document version 1.2, 25 March 2026. https://www.gleif.org/organizational-identity/become-a-vlei-issuer-qvi/vlei-ecosystem-governance-framework/2026-03-25_vlei-egf-v4.0-primary-document_v1.2_final.pdf (visited on Oct. 21, 2026).

Google and contributors (2025). *Agent Payments Protocol (AP2): Specification*. https://ap2-protocol.org/specification/ (visited on Sept. 4, 2026).

Google, Shopify and the UCP co-developers (2026). *Universal Commerce Protocol (UCP): Specification*. Announced 11 January 2026; Apache 2.0; capabilities Catalog, Cart, Checkout, Identity Linking, Order. https://ucp.dev/ (visited on Oct. 22, 2026).

Gudgeon, L., P. Moreno-Sanchez, S. Roos, P. McCorry, and A. Gervais (2020). "SoK: Layer-Two Blockchain Protocols". In: *Financial Cryptography and Data Security (FC 2020)*. Vol. 12059. Lecture Notes in Computer Science. Springer, pp. 201–226. DOI: 10.1007/978-3-030-51280-4_12. https://eprint.iacr.org/2019/360 (visited on Sept. 4, 2026).

Hardy, N. (1988). "The Confused Deputy (or why capabilities might have been invented)". In: *ACM SIGOPS Operating Systems Review* 22.4, pp. 36–38. DOI: 10.1145/54289.871709.

Hu, B. and H. Rong (2025). *Inter-Agent Trust Models: A Comparative Study of Brief, Claim, Proof, Stake, Reputation and Constraint in Agentic Web Protocol Design — A2A, AP2, ERC-8004, and Beyond*. arXiv: 2511.03434.

Kamvar, S. D., M. T. Schlosser, and H. Garcia-Molina (2003). "The EigenTrust Algorithm for Reputation Management in P2P Networks". In: *Proceedings of the 12th International Conference on World Wide Web (WWW)*, pp. 640–651. DOI: 10.1145/775152.775242.

Kim, P. J., K. Britz, and D. Knott (2020). *EIP-3009: Transfer With Authorization*. Ethereum Improvement Proposal, Standards Track (ERC), Draft; created 28 September 2020. https://eips.ethereum.org/EIPS/eip-3009 (visited on Oct. 21, 2026).

Komlo, C. and I. Goldberg (2021). "FROST: Flexible Round-Optimized Schnorr Threshold Signatures". In: *Selected Areas in Cryptography – SAC 2020*. Vol. 12804. Lecture Notes in Computer Science. Also: Cryptology ePrint Archive, Paper 2020/852. Springer, pp. 34–65. https://eprint.iacr.org/2020/852 (visited on Sept. 4, 2026).

Lampson, B., M. Abadi, M. Burrows, and E. Wobber (1992). "Authentication in Distributed Systems: Theory and Practice". In: *ACM Transactions on Computer Systems* 10.4, pp. 265–310. DOI: 10.1145/138873.138874.

Lan, Q., A. Kaul, S. Jones, and S. Westrum (2026). *Zero-Trust Runtime Verification for Agentic Payment Protocols: Mitigating Replay and Context-Binding Failures in AP2*. arXiv: 2602.06345.

Laurie, B., A. Langley, and E. Kasper (2013). *Certificate Transparency (RFC 6962)*. https://www.rfc-editor.org/rfc/rfc6962 (visited on Sept. 4, 2026).

Li, Y., L. Wang, K. Wang, Z. Yang, K. Wang, Z. Guan, and J. Gao (2026a). *A402: Binding Cryptocurrency Payments to Service Execution for Agentic Commerce*. arXiv: 2603.01179.

Li, Z., Q. Wang, and Z. Wang (2026b). *Five Attacks on x402 Agentic Payment Protocol*. arXiv: 2605.11781.

Ling, S., Y. Huang, Y. Du, Y. Chen, Y. Zhou, L. Wu, and C. Wang (2026). *Free-Riding the Agentic Web: A Systematic Security Analysis of x402 Payments*. arXiv: 2605.30998.

Mastercard (2025). *Scaling Agentic Commerce with Trust*. 14 October 2025. Mastercard Agent Pay, announced 29 April 2025; merchant specifications (Agent Pay Acceptance Framework) on Mastercard Developers. https://www.mastercard.com/global/en/news-and-trends/stories/2025/lady-gaga-dance-contest-abracadabra-fan-version/scaling-agentic-commerce-with-trust.html (visited on Oct. 22, 2026).

McWhirter, S. and D. Montes Serralde (2026). *Know Your Agent: Building the Compliance Layer for Agentic Finance*. Binance Global Policy Insights, October 2026. https://www.binance.com/en/square/post/372666808420252 (visited on Oct. 22, 2026).

Meiklejohn, S., M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage (2013). "A Fistful of Bitcoins: Characterizing Payments among Men with No Names". In: *Proceedings of the 2013 Internet Measurement Conference (IMC)*. ACM, pp. 127–140. DOI: 10.1145/2504730.2504747. https://cseweb.ucsd.edu/~smeiklejohn/files/imc13.pdf (visited on Oct. 22, 2026).

Ménétrey, J., C. Göttel, M. Pasin, P. Felber, and V. Schiavoni (2022). "An Exploratory Study of Attestation Mechanisms for Trusted Execution Environments". In: *Workshop on System Software for Trusted Execution (SysTEX)*. arXiv: 2204.06790.

Miller, M. S., K.-P. Yee, and J. Shapiro (2003). *Capability Myths Demolished*. Tech. rep. SRL2003-02. Johns Hopkins University Systems Research Laboratory. http://zesty.ca/capmyths/ (visited on Sept. 4, 2026).

Murdoch, S. J. and R. Anderson (2014). "Security Protocols and Evidence: Where Many Payment Systems Fail". In: *Financial Cryptography and Data Security (FC 2014)*. Vol. 8437. Lecture Notes in Computer Science. Springer, pp. 21–32. https://discovery-pp.ucl.ac.uk/id/eprint/10074689 (visited on Sept. 4, 2026).

Nevermined (2026). *x402 Delegation Extension*. https://nevermined.ai/docs/specs/x402-card-delegation (visited on Sept. 4, 2026).

OpenAI and Stripe (2026). *Agentic Commerce Protocol (ACP): Specification*. Specification version 2026-04-17 (beta), Apache 2.0; checkout and delegate-payment specifications. Stripe's Shared Payment Token reference: https://docs.stripe.com/api/shared-payment/issued-token/object. https://github.com/agentic-commerce-protocol/agentic-commerce-protocol (visited on Oct. 22, 2026).

Pandy, S. and M. Crowe (2019). *Industry Perspectives on the Evolution of EMV Payment Tokenization*. White paper. Revised 6 May 2019; originally released 24 September 2018. Federal Reserve Bank of Boston, Payments Strategies Group. https://www.bostonfed.org/-/media/Documents/PaymentStrategies/evolution-of-payment-tokenization.pdf (visited on Oct. 21, 2026).

Pass, R. and a. shelat abhi (2015). "Micropayments for Decentralized Currencies". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Also: Cryptology ePrint Archive, Paper 2016/332, pp. 207–218. DOI: 10.1145/2810103.2813713. https://eprint.iacr.org/2016/332 (visited on Sept. 4, 2026).

Poon, J. and T. Dryja (2016). *The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments*. https://lightning.network/lightning-network-paper.pdf (visited on Sept. 4, 2026).

Proos, J. and C. Zalka (2003). "Shor's Discrete Logarithm Quantum Algorithm for Elliptic Curves". In: *Quantum Information and Computation* 3.4, pp. 317–344. DOI: 10.26421/QIC3.4-3. arXiv: quant-ph/0301141. https://arxiv.org/abs/quant-ph/0301141 (visited on Nov. 11, 2026).

Resnick, P. and R. Zeckhauser (2002). "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System". In: *The Economics of the Internet and E-Commerce*. Ed. by M. R. Baye. Vol. 11. Advances in Applied Microeconomics. Emerald Group Publishing, pp. 127–157. DOI: 10.1016/S0278-0984(02)11030-3. https://scholar.harvard.edu/sites/scholar.harvard.edu/files/rzeckhauser/files/trust_among_strangers.pdf (visited on Sept. 4, 2026).

Rivest, R. L. (1998). "Can We Eliminate Certificate Revocation Lists?" In: *Financial Cryptography (FC '98)*. Ed. by R. Hirschfeld. Vol. 1465. Lecture Notes in Computer Science. Springer, pp. 178–183. DOI: 10.1007/BFb0055482. https://people.csail.mit.edu/rivest/pubs/Riv98b.pdf (visited on Oct. 21, 2026).

Rivest, R. L. and B. Lampson (1996). *SDSI — A Simple Distributed Security Infrastructure*. Working document, version 1.1. https://people.csail.mit.edu/rivest/pubs/RL96.ver-1.1.html (visited on Sept. 4, 2026).

Rivest, R. L. and A. Shamir (1997). "PayWord and MicroMint: Two Simple Micropayment Schemes". In: *Security Protocols Workshop 1996*. Vol. 1189. Lecture Notes in Computer Science. Springer, pp. 69–87. DOI: 10.1007/3-540-62494-5_6. https://people.csail.mit.edu/rivest/pubs/RS96a.pdf (visited on Sept. 4, 2026).

Russinovich, M., M. Costa, C. Fournet, D. Chisnall, A. Delignat-Lavaud, S. Clebsch, K. Vaswani, and V. Bhatia (2021). "Toward Confidential Cloud Computing". In: *ACM Queue* 19.1, pp. 49–76. DOI: 10.1145/3454122.3456125. https://queue.acm.org/detail.cfm?id=3456125 (visited on Oct. 22, 2026).

Saltzer, J. H. and M. D. Schroeder (1975). "The Protection of Information in Computer Systems". In: *Proceedings of the IEEE* 63.9, pp. 1278–1308. DOI: 10.1109/PROC.1975.9939. https://web.mit.edu/Saltzer/www/publications/protection/ (visited on Oct. 21, 2026).

See, K. and X. W. Tan (2026). *Compliance-Aware Agentic Payments on Stablecoin Rails*. arXiv: 2605.00071.

Shor, P. W. (1997). "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer". In: *SIAM Journal on Computing* 26.5, pp. 1484–1509. DOI: 10.1137/S0097539795293172. arXiv: quant-ph/9508027. https://arxiv.org/abs/quant-ph/9508027 (visited on Nov. 11, 2026).

South, T., S. Marro, T. Hardjono, R. Mahari, C. D. Whitney, A. Chan, and A. Pentland (2025). "Position: AI Agents Need Authenticated Delegation". In: *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*. Vol. 267. Proceedings of Machine Learning Research. Extended version: arXiv:2501.09674. PMLR, pp. 82211–82231. https://proceedings.mlr.press/v267/south25a.html (visited on Oct. 21, 2026).

Tempo Labs and Stripe (2026). *Machine Payments Protocol (MPP) Specifications*. Core specification also as IETF Internet-Draft draft-ryan-httpauth-payment-01, *The "Payment" HTTP Authentication Scheme* (Ryan, Moxey, Meagher, Weinstein, Kaliski), 17 March 2026, individual submission. https://github.com/tempoxyz/mpp-specs (visited on Oct. 21, 2026).

Tessaro, S. and C. Zhu (2023). "Revisiting BBS Signatures". In: *Advances in Cryptology – EUROCRYPT 2023*. Vol. 14008. Lecture Notes in Computer Science. Also: Cryptology ePrint Archive, Paper 2023/275. Springer, pp. 691–721. DOI: 10.1007/978-3-031-30589-4_24. https://eprint.iacr.org/2023/275 (visited on Sept. 4, 2026).

UCAN Working Group (2023). *UCAN: User-Controlled Authorization Networks Specification*. https://github.com/ucan-wg/spec (visited on Sept. 4, 2026).

United States Congress (2025). *Guiding and Establishing National Innovation for U.S. Stablecoins Act (GENIUS Act)*. Public Law 119–27, 18 July 2025, 139 Stat. 419; Sections 2(16) and 4(a)(6)(B). https://www.govinfo.gov/content/pkg/PLAW-119publ27/pdf/PLAW-119publ27.pdf (visited on Oct. 22, 2026).

Visa (2025). *Trusted Agent Protocol: Specifications*. Merchant specification, Visa Developer Platform; protocol announced 14 October 2025. https://developer.visa.com/docs/trusted-agent-protocol/trusted-agent-protocol-specifications (visited on Oct. 22, 2026).

Visa (2026a). *Visa Core Rules and Visa Product and Service Rules*. Public edition of 18 April 2026; Section 11, Dispute Resolution, and the dispute conditions of categories 10–13. https://usa.visa.com/dam/VCOM/download/about-visa/visa-rules-public.pdf (visited on Oct. 22, 2026).

Visa (2026b). *Visa Dispute Management Guide for Merchants: Best Practices, Tools and Resources*. October 2026 edition. https://usa.visa.com/dam/VCOM/global/support-legal/documents/merchants-dispute-management-guidelines.pdf (visited on Oct. 22, 2026).

W3C (2025). *Verifiable Credentials Data Model v2.0*. W3C Recommendation, 15 May 2025; with Verifiable Credential Data Integrity 1.0, W3C Recommendation of the same date. https://www.w3.org/TR/vc-data-model-2.0/ (visited on Oct. 22, 2026).

Walfish, M. and A. J. Blumberg (2015). "Verifying Computations without Reexecuting Them". In: *Communications of the ACM* 58.2, pp. 74–84. DOI: 10.1145/2641562.

Waller, C. J. (2026). *Payments in the Age of AI Agents*. Speech, Sibos 2026, Miami, 29 September 2026. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/newsevents/speech/files/waller20260929a.pdf (visited on Oct. 21, 2026).

WE BUILD Consortium (2026). *Trusted Identities for AI Agents: An Opportunity for Europe*. Non-paper, 27 February 2026. https://www.webuildconsortium.eu/news/non-paper-trusted-identities-for-ai-agents-an-opportunity-for-europe-8396 (visited on Oct. 21, 2026).

World Foundation (2026). *World Whitepaper*. Four-part series, updated 25 March 2026. https://whitepaper.world.org/ (visited on Oct. 22, 2026).

Xiong, X., Z. Li, W. Wei, Q. Wang, W. Knottenbelt, and Z. Wang (2026). *Can Trustless Agents Be Trusted? An Empirical Study of the ERC-8004 Decentralized AI Agent Ecosystem*. arXiv: 2606.26028.

Zhang, Y., Y. Xiang, Y. Lei, Q. Wang, T. Qiu, Y. Sun, et al. (2026). *SoK: Blockchain Agent-to-Agent Payments*. arXiv: 2604.03733.