READING GROUP SYLLABUS

# Cryptography for Agentic Commerce

Ideas and techniques for making agentic commerce trustworthy with cryptography

Christian Pfrang | 4 September 2026

**Christian Pfrang** Markets, Policy and Technology

Draft for discussion

# 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!").

This reading group discusses various types of assurances that buyers and sellers might want about each other and regarding the faithfulness of their agents’ actions. We will explore papers and writings from cryptography, security engineering and the emergent literature in agentic commerce, including agent commerce protocols.

The goal is to develop a better understanding of the mechanisms necessary to integrate AI agents into commercial decision making, to gain and intuition about attack surfaces and failure modes, and to identify cryptographic mechanisms that might be useful in providing these assurances.

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

In this session we consider the most vivid structural failure mode for delegated agents and use it to review and develop key mental models for delegated authority in automated systems.

What if an agent, for example through a **prompt injection** attack, acts differently than what the principal that delegated it had intended it to?

The agent under attack acts as a **confused deputy**, because it acts under the authority of the principal (“send the money") but using the instructions of the attacker (“to my bank account").

**“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."

We will then look at how these mental models are reflected in modern **agent-to-agent commerce protocols.**

## 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/
- **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
- **SoK: Blockchain Agent-to-Agent Payments** — Zhang, Y. et al., 2026 — Four-stage lifecycle (discovery, authorization, execution, accounting) — arxiv.org/abs/2604.03733

## 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/
- **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

# 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.

**“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.

An important question to consider 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.

## 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
- **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.

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.

### 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

### Artifacts

- **x401 Specification** — x401 contributors, 2026 — Proposed agent-authentication counterpart to x402 on the HTTP 401 side. — https://github.com/proof/x401
- **ERC-8004: Trustless Agents** — Ethereum community, 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

## 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.

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

### Artifacts

- **x402 Protocol Specification** — Coinbase, 2025 — The protocol spec itself: header, facilitator API, payment schemes. — https://github.com/coinbase/x402
- **EIP-3009: Transfer With Authorization** — Coinbase, 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, 2025 — The "Payment" HTTP authentication scheme; the main contender next to x402. — https://github.com/tempoxyz/mpp-specs

## 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?

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 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 — Pair with the previous item: breaks the fix. — 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

## 6 Session 6: Selective disclosure and compliance predicates

In session 2 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.

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.

**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.

### 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

### Artifacts

- **European Digital Identity Wallet: Architecture and Reference Framework** — European Commission, 2025 — The EU wallet architecture; its zero-knowledge-proof section sketches where selective disclosure lands in regulation. — https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/
- **Selective Disclosure for JWTs (SD-JWT)** — Fett, D. et al., 2026 — Selective disclosure by salted hashes over JWT claims — the pragmatic, linkable baseline. — https://datatracker.ietf.org/doc/draft-ietf-oauth-selective-disclosure-jwt/

## 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.

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.

## 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
- **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

## Artifacts

- **Pay with Confidence: Verifiable, Auditable Agent Payments on Amazon Bedrock AgentCore** — Solv Labs and AWS, 2026 — The design under discussion: zero-knowledge proof of policy evaluation plus TEE attestation for agent payments. — https://aws.amazon.com/blogs/machine-learning/pay-with-confidence-how-solv-labs-built-verifiable-auditable-agent-payments-on-amazon-bedrock-agentcore-payments/
- **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. — https://www.rfc-editor.org/rfc/rfc6962

# 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?

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 using without a central party managing the counter state.

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. Sets up the World discussion without presupposing its answer. — 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. — https://eprint.iacr.org/2006/454
- **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, 2023 — Proof of human at scale: iris biometrics, the Orb, and World ID. — 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.

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).

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).

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).

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).

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).

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. 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).

Coinbase (2020). *EIP-3009: Transfer With Authorization*. https://eips.ethereum.org/EIPS/eip-3009 (visited on Sept. 4, 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.

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.

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).

Ethereum community (2025). *ERC-8004: Trustless Agents*. https://eips.ethereum.org/EIPS/eip-8004 (visited on Sept. 4, 2026).

European Commission (2025). *European Digital Identity Wallet: Architecture and Reference Framework*. https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/ (visited on Sept. 4, 2026).

Fett, D., K. Yasuda, and B. Campbell (2026). *Selective Disclosure for JWTs (SD-JWT)*. IETF Internet-Draft. https://datatracker.ietf.org/doc/draft-ietf-oauth-selective-disclosure-jwt/ (visited on Sept. 4, 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.

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).

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

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.

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).

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).

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. 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).

Solv Labs and AWS (2026). *Pay with Confidence: Verifiable, Auditable Agent Payments on Amazon Bedrock AgentCore*. 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 Sept. 4, 2026).

Tempo (2025). *Machine Payments Protocol (MPP) Specifications*. https://github.com/tempoxyz/mpp-specs (visited on Sept. 4, 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).

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.

World Foundation (2023). *World Whitepaper*. https://whitepaper.world.org/ (visited on Sept. 4, 2026).

x401 contributors (2026). *x401 Specification*. https://github.com/proof/x401 (visited on Sept. 4, 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.