Computer Fraud and Security
ISSN (online): 1873-7056

Zero-Knowledge Mandates: Privacy-Preserving Delegation & Spend Controls for AP2 Across Heterogeneous Rails

Hirenkumar Patel

Mastercard Inc, USA

Abstract

Current agent payment standards enable transactions across varied infrastructure, including card systems, banking channels, and blockchain platforms, through cryptographic mandates binding user intentions to agent actions. These mandates create authorization structures while revealing critical vulnerabilities in transaction privacy protection, fine-grained delegation management, and cohesive governance implementation across multiple payment infrastructures. Zero-Knowledge Mandates introduce cryptographic techniques allowing agents to demonstrate compliance with spending restrictions while concealing constraint details from verifiers. Agents demonstrate compliance with spending caps, approved vendors, and time restrictions while keeping financial details and payment channel choices hidden. The system uses compact cryptographic proofs that allow verification without exposing mandate terms, user account information, or transaction routing.

Core security guarantees include execution unlinkability, preventing transaction correlation, and verifiable compliance, ensuring constraint adherence. Technical implementation utilizes efficient proof systems, maintaining real-time transaction processing requirements. Evaluation addresses computational performance, information leakage boundaries, and practical deployment considerations across heterogeneous payment networks. The resulting architecture provides the first comprehensive privacy-preserving authorization primitive for autonomous commercial agents operating across multiple financial infrastructures simultaneously.

Keywords: Agentic commerce · Agent Payments Protocol (AP2), Zero-knowledge proofs, Delegation control, Privacy-preserving payments, Heterogeneous payment rails, Verifiable credentials, Spend caps · Auditability.

1. Introduction: The Privacy Gap in Agentic Commerce

The integration of autonomous agents into financial operations has created urgent demands for secure delegation frameworks that balance authorization with privacy. Current protocols establish cryptographically-signed mandates binding user intentions to agent actions, yet critical privacy deficiencies persist [1]. Autonomous agents must execute transactions on behalf of users while maintaining clear authorization boundaries and accountability chains across diverse payment infrastructures. The Agent Payments Protocol represents a significant advancement in standardizing these delegation mechanisms through cryptographic mandates, yet fundamental privacy gaps remain unaddressed [2]. This framework addresses how autonomous agents can prove transaction compliance with user-defined constraints without revealing the constraints themselves or underlying financial data. The challenge lies in achieving verifiable compliance while preserving transactional privacy across heterogeneous payment systems, including card networks, banking channels, and digital asset platforms.

1.1 Agentic Commerce and Delegation Requirements

Autonomous agents powered by advanced language models are increasingly executing transactions across financial and enterprise domains, fundamentally transforming commercial interactions. These agents require standardized delegation mechanisms that maintain clear authorization boundaries while enabling flexible operation across diverse transaction contexts [3]. The deployment acceleration of these systems necessitates robust frameworks capable of managing complex authorization scenarios where users delegate specific purchasing powers to agents operating with varying degrees of autonomy. The Agent Payments Protocol introduces verifiable mandates—Intent, Cart, and Payment types—anchored in cryptographic signatures to establish agent authorization. These mandates create verifiable credentials linking user intentions to agent actions across diverse commercial contexts, providing a foundational trust layer for agentic commerce.

However, current delegation models present significant limitations in granularity and privacy preservation [4]. Enabling broad delegation capabilities, such as authorizing purchases below certain thresholds during specific timeframes or within particular merchant categories, requires mechanisms that preserve constraint confidentiality while enabling verification. Users need delegation frameworks that grant agents operational flexibility without exposing detailed spending rules, authorized merchant lists, or temporal restrictions to verifying parties. The requirement for flexible yet private delegation becomes particularly acute in enterprise contexts where procurement agents must operate under complex policy constraints while maintaining competitive confidentiality. Traditional mandate structures force a binary choice between transparency and functionality, where either all constraint details are exposed for verification or delegation capabilities remain severely limited. This limitation hinders the practical deployment of autonomous agents in sensitive commercial environments where both operational flexibility and privacy preservation are non-negotiable requirements.

1.2 Privacy and Control Gaps in Agent Payments Protocol

Current mandate verification mechanisms create substantial privacy vulnerabilities that threaten user confidentiality and commercial sensitivity. The protocol relies on traditional signature schemes and third-party auditors, requiring merchants and payment infrastructure to access mandated terms directly [5]. This verification architecture necessitates exposing transaction details and linking activities to broader user financial profiles, creating comprehensive data trails vulnerable to aggregation and analysis.

2. Problem Statement and Requirements

The framework must address authorization verification, constraint enforcement, cross-infrastructure compatibility, dispute resolution, and privacy preservation simultaneously [7]. Satisfying all requirements simultaneously demands novel cryptographic approaches that enable selective disclosure while maintaining verifiable accountability across heterogeneous payment ecosystems [8].

2.1 Delegation and Spend Control Requirements

Authorization verification presents the foundational challenge: agents must prove possession of valid delegation authority under specific constraints without revealing constraint details or user identities. Delegation proofs must demonstrate that user-granted authority encompasses the proposed transaction under parameters including spending caps, temporal windows, and categorical restrictions [9]. However, exposing these parameters to verifying parties creates privacy leakage and potential security vulnerabilities. A user delegating authority for office supply purchases up to weekly limits should not reveal the specific cap amount, remaining budget, or historical spending patterns to merchants or payment processors.

Spend control enforcement introduces additional complexity at transaction execution. The system must cryptographically prove that proposed transactions satisfy mandated constraints—spending amounts, merchant categories, temporal validity—without granting verifiers access to complete constraint specifications or user financial states [10]. Traditional verification approaches require exposing either the full constraint set or individual constraint values, forcing unnecessary information disclosure. A merchant processing a delegated transaction needs confirmation of authorization and constraint compliance, but requires no visibility into the user's total budget, other authorized categories, or the agent's internal decision logic. The challenge intensifies when constraints involve complex predicates combining multiple conditions, such as categorical spending limits that reset periodically or merchant whitelists that vary by transaction context. Cryptographic proofs must accommodate these sophisticated constraint structures while preserving privacy across all constraint dimensions simultaneously.

2.2 Rail-Agnostic Verification and Auditability

Cross-infrastructure compatibility demands that verification mechanisms function consistently across diverse payment systems despite fundamental differences in transaction semantics, audit requirements, and settlement processes [11]. Card networks operate under different disclosure frameworks than bank transfer systems, which differ substantially from blockchain-based digital asset platforms. Each infrastructure maintains distinct requirements for transaction validation, fraud prevention, and regulatory compliance. Delegation proofs must accommodate these varied requirements without fragmenting privacy guarantees or requiring rail-specific customization that increases implementation complexity and reduces interoperability.

Auditability requirements further complicate privacy preservation objectives. Regulatory frameworks and dispute resolution processes demand transparent attribution chains linking users, agents, merchants, and transactions [12]. When fraud occurs or disputes arise, relevant parties must reconstruct authorization chains to establish accountability and assign responsibility. However, routine transaction verification should not grant broad access to these attribution chains. The framework must support selective auditability where authorized parties can access necessary information under specific circumstances while maintaining transactional privacy during normal operations. This selective disclosure extends to regulatory compliance, where supervisory authorities require audit capabilities without compromising user privacy during standard transaction processing. Privacy minimization principles demand limiting disclosure of personally identifiable information, spending patterns, agent operational details, and merchant transaction data beyond strict verification and settlement requirements. Achieving this balance requires cryptographic mechanisms enabling granular control over information exposure calibrated to specific verification contexts and authorization levels.

3. Zero-Knowledge Mandate Framework

The Zero-Knowledge Mandate Framework establishes a cryptographic layer extending existing payment protocols to enable privacy-preserving delegation verification. The framework introduces cryptographic primitives allowing agents to demonstrate transaction compliance with user-defined constraints without exposing constraint specifications, financial details, or payment infrastructure selections. By embedding succinct non-interactive arguments into mandate structures, the system enables verification parties to confirm authorization validity and constraint satisfaction while learning nothing beyond proof validity itself. This approach fundamentally transforms delegation verification from information-exposing processes into privacy-preserving cryptographic protocols, maintaining accountability without compromising confidentiality across heterogeneous payment environments [2].

3.1 ZK-Mandate Structure and Definition

Zero-Knowledge Mandates are defined as cryptographic tuples containing mandate identifiers, proofs, and public parameters enabling verification without knowledge disclosure. The structure extends traditional mandate types—Intent, Cart, and Payment—used in agent payment protocols [3]. The framework embeds verifiable credentials, incorporating cryptographic commitments and zero-knowledge proofs within these mandated structures. Users sign credentials containing commitments to constraint parameters—spending caps, authorized categories, and validity windows—alongside proofs demonstrating commitment satisfaction of delegation policies [4]. Agents present these credentials to merchants and issuers during transactions, enabling verification without revealing underlying constraint values or user financial states. The cryptographic construction employs succinct non-interactive arguments of knowledge, specifically zk-SNARKs or zk-STARKs, generating proofs demonstrating two fundamental properties: authority possession confirming user authorization for mandate execution, and constraint compliance proving proposed transactions satisfy all policy conditions within mandate specifications. Mandate details and private inputs remain confidential to agents and users while generated proofs remain publicly verifiable.

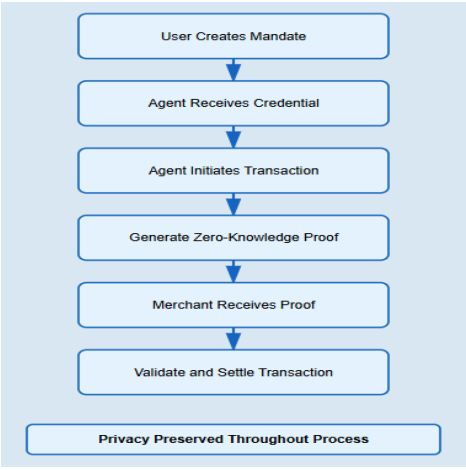

Figure 1: Zero-Knowledge Mandate Transaction Flow [3], [4]

Mandate Component	Function
Intent Mandate	Establishes delegation parameters, including spending limits, merchant categories, and temporal boundaries
Cart Mandate	Captures transaction details, including itemization, pricing, and fulfillment specifications
Payment Mandate	Signals agent involvement to payment networks, enabling risk assessment and authorization routing
Cryptographic Commitment	Binds constraint parameters without revealing underlying values
Zero-Knowledge Proof	Demonstrates constraint satisfaction while maintaining confidentiality
Verifiable Credential	Enables verification without exposing financial states or constraint details

Table 1: Mandate Types and Functions [3], [4]

3.2 Spend Controls and Policy Enforcement

Policy enforcement mechanisms utilize domain-specific languages defining spend controls within mandate specifications. These languages are designed for efficient compilation into arithmetic circuits suitable for zero-knowledge proof generation, specifically Rank-1 Constraint Systems enabling cryptographic verification [5]. The framework enforces multiple constraint categories through zero-knowledge proofs, maintaining privacy across all dimensions simultaneously. Balance constraints prove that spending amounts satisfy minimum balance requirements without revealing actual balance values or transaction amounts. Rate limit constraints demonstrate that cumulative spending, including current transactions, remains within periodic caps without exposing individual transaction histories or cap values.

Merchant authorization constraints prove transaction recipients appear in authorized lists without revealing complete authorization sets [6]. Temporal validity constraints confirm that delegation remains active within specified time windows without exposing window boundaries or usage histories. Category restrictions prove transactions fall within delegated purchasing categories without revealing full category specifications or alternative authorized categories. Each constraint type employs cryptographic techniques enabling verification parties to confirm satisfaction without accessing underlying constraint parameters or financial data. Proof frameworks, including Bulletproofs, zk-SNARKs, and zk-STARKs, provide the cryptographic foundation for these constraint verifications, offering different trade-offs between proof size, generation time, and verification efficiency suitable for varied deployment contexts across payment infrastructures.

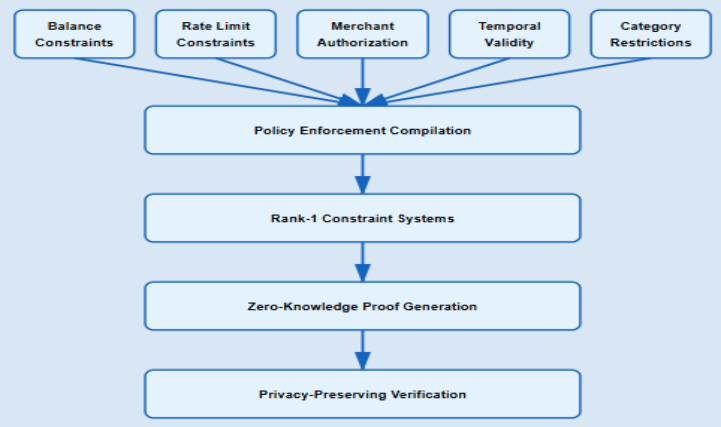

Figure 2: Constraint Verification Categories [5], [6]

4. Cross-Rail Implementation and Security Architecture

Cross-infrastructure deployment requires standardized verification mechanisms functioning consistently across diverse payment systems while maintaining uniform security guarantees. The framework introduces a rail abstraction layer enabling mandate verification across card networks, bank transfer systems, and digital asset platforms without requiring infrastructure-specific customization [7]. Verification oracles deployed at consumption points—merchants or payment rails—validate zero-knowledge proofs using only succinct proof artifacts and public transaction parameters such as destination accounts or wallet addresses. Rails confirms compliance based solely on cryptographic proof validity, abstracting underlying policy mechanisms and constraint specifications. This architecture represents the first unified approach for privacy-preserving spend controls across heterogeneous financial backends operating under distinct disclosure requirements and audit frameworks [8].

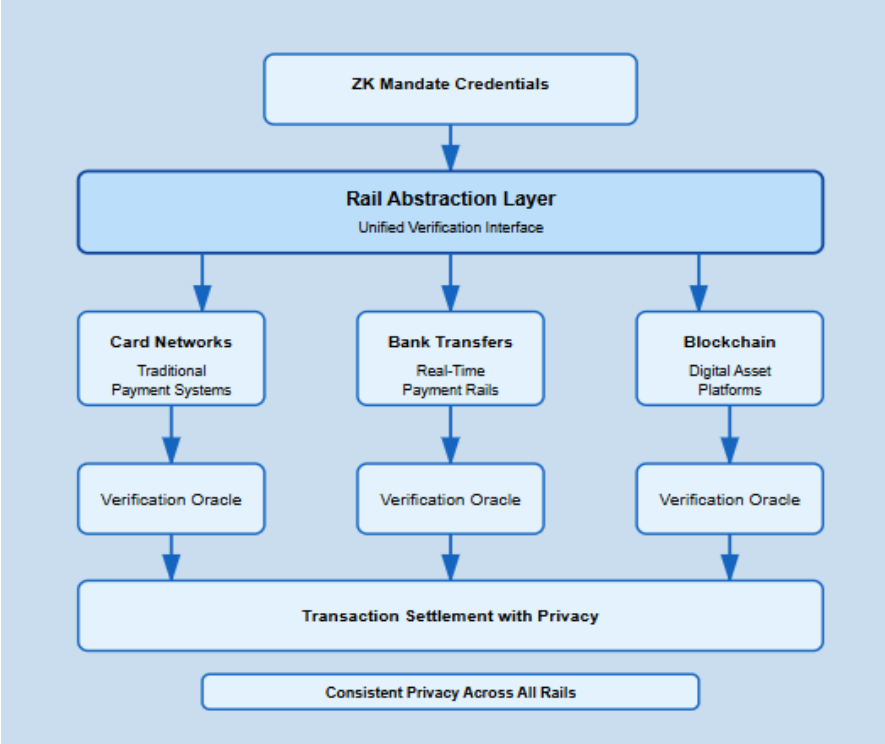

Figure 3: Cross-Rail Architecture [7], [8]

The Mandate Evidence Record provides a standardized metadata structure captured within each rail's transaction payload, containing cryptographic hashes referencing zero-knowledge mandate credentials and proof metadata. Regardless of payment rail selection—card network, real-time payment system, or blockchain platform—verifying parties can validate mandate compliance without custom rail-specific verification logic. The formal security model defines participating entities, including users, agents, merchants, and payment networks, alongside mandate constraint structures specifying caps, category sets, time windows, and usage limits [9]. Commitment schemes bind constraint values cryptographically, while zero-knowledge proofs demonstrate transaction validation against committed constraints.

Component	Implementation Details
Verification Oracle	Validates zero-knowledge proofs at consumption points using succinct proof artifacts and public transaction parameters
Mandate Evidence Record	Standardized metadata structure containing cryptographic hashes referencing credentials and proof metadata across rails
Rail Abstraction Layer	Enables mandate verification across card networks, bank transfers, and digital assets without infrastructure-specific customization

Commitment Scheme	Cryptographically binds constraint values while enabling zero-knowledge proof generation for transaction validation
Audit Trail	Maintains linkage between mandate identifiers, proof identifiers, and transaction identifiers, enabling dispute resolution
Policy Mechanism	Abstracts underlying constraint specifications allowing rails to confirm compliance through cryptographic proof validity

Table 2: Cross-Rail Implementation Components [7], [8]

Security properties include authorization correctness, ensuring only agents possessing valid user-signed mandates can initiate transactions, spend control enforcement, preventing transactions exceeding constraints without new mandate issuance, and rail-agnostic accountability, enabling any verifier to link transactions to mandates through proofs [10]. Privacy guarantees ensure verifiers learn only proof validity without inferring user budgets, spending histories, or complete constraint specifications. Threat model analysis addresses agent compromise, replay attacks, mandate reuse, and collusion scenarios, demonstrating how zero-knowledge proofs prevent abuse, including attempts to exceed spending caps through invalid proof generation.

5. Implementation, Evaluation, and Security Validation

The framework satisfies four fundamental security properties essential for financial delegation systems. Zero-knowledge privacy guarantees ensure verifiers—merchants and payment rails—learn nothing about mandate specifications or private inputs beyond constraint satisfaction confirmation [11]. Soundness properties prevent dishonest agents from generating valid proofs for non-compliant transactions with non-negligible probability, maintaining enforcement integrity through cryptographic hardness assumptions. Unlinkability of execution prevents correlation of proof sequences to specific users by verifiers or external observers unless explicitly required for regulatory auditing or dispute resolution scenarios. Limited non-repudiation maintains user accountability by preventing mandate issuance denial while preserving mandate confidentiality through cryptographic commitments signed by users [12].

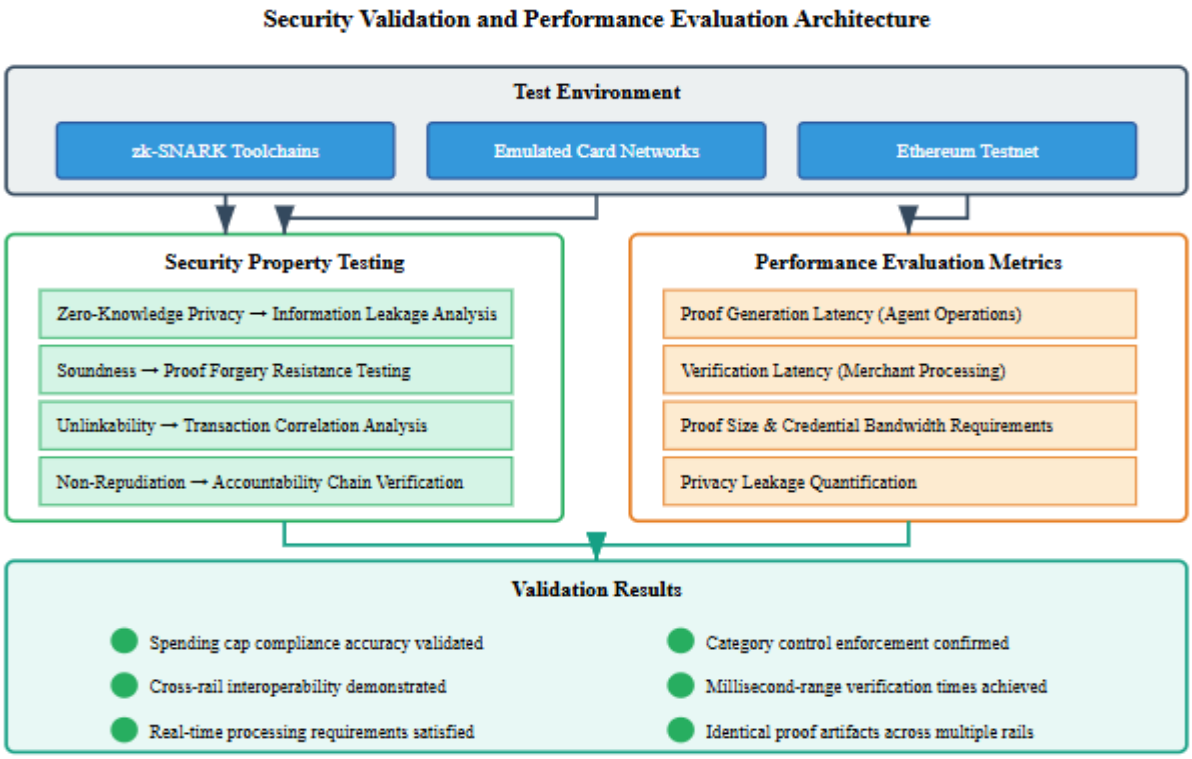

Figure 4: Security Validation and Performance Evaluation Architecture [11]

Proof-of-concept implementation employs zk-SNARK toolchains across emulated card networks and Ethereum testnet stablecoin deployments, validating cross-rail verification capabilities [11]. Performance evaluation addresses multiple critical metrics, including proof generation latency impacting agent operations and verification latency affecting merchant transaction processing. Proof and credential size measurements assess bandwidth requirements and payload impacts on transaction throughput. Privacy leakage quantification provides formal measures of information accessible to adversaries through observation of proof artifacts and transaction patterns.

Evaluation confirms spending cap compliance and category control accuracy across varied configurations. Cross-rail interoperability validation demonstrates identical proof artifacts functioning across payment infrastructures with minimal adjustment. Modern succinct zero-knowledge proof schemes optimized for arithmetic constraints typical of financial logic—including addition and comparison operations—maintain verification times in millisecond ranges, satisfying real-time requirements for commercial agent transactions despite proof generation representing the primary computational bottleneck.

Security Property	Description
Zero-Knowledge Privacy	Verifiers learn nothing about mandate specifications or private inputs beyond constraint satisfaction confirmation
Soundness	Dishonest agents cannot generate valid proofs for non-compliant transactions with cryptographic hardness guarantees
Unlinkability	Proof sequences cannot be correlated to specific users by verifiers or external observers during routine operations
Limited Non-Repudiation	Users cannot deny mandate issuance while maintaining mandate confidentiality through cryptographic commitments
Authorization Correctness	Only agents with valid user-signed mandates can initiate transactions within delegated boundaries
Constraint Enforcement	Transactions exceeding mandate constraints require new user authorization, preventing unauthorized spending

Table 3: ZK-Mandate Security Properties [11]

6. Implications and Future Directions

Zero-knowledge mandates introduce significant implications across payment industry ecosystems, regulatory frameworks, and agentic commerce deployment contexts. For payment networks and issuers, the framework enhances user privacy and delegation flexibility while providing structured, verifiable control mechanisms essential as agentic commerce scales to mainstream adoption. Networks gain cryptographic assurance of transaction authorization and constraint compliance without accessing sensitive user data or detailed spending patterns, reducing liability while maintaining security standards. Regulatory and compliance frameworks benefit from audit-ready evidence chains supporting minimal disclosure principles aligned with data protection regulations, including privacy standards and payment card industry requirements. The architecture enables strong customer authentication similar to regulatory frameworks while preserving transactional privacy through selective disclosure mechanisms.

The agentic commerce ecosystem gains flexible delegation models supporting complex scenarios, including recurring budget allocations and categorical spending authorities, without exposing complete spending histories. This capability enables new deployment contexts such as procurement automation and corporate agent operations requiring sophisticated authorization structures. Current limitations include zero-knowledge proof computational overhead, standardization challenges across diverse payment infrastructures, legacy rail integration friction, and user education requirements for agent trust models.

Future development directions encompass dynamic delegation supporting agent renegotiation of constraints, multi-agent delegation hierarchies enabling organizational structures, richer policy languages accommodating complex compliance requirements, and incorporation of real-time risk assessment signals into zero-knowledge proofs. Circuit optimization for

sophisticated policy expressions, hierarchical delegation supporting enterprise scenarios, and adaptive risk evaluation integration represent priority enhancement areas. These extensions position zero-knowledge mandates as foundational infrastructure for privacy-conscious autonomous commerce operating under stringent confidentiality requirements within regulated financial environments. Standardization efforts across payment networks will accelerate adoption while ensuring interoperability.

Metric Category	Evaluation Focus
Proof Generation Latency	Computational time required for agents to generate zero-knowledge proofs, impacting transaction initiation speed
Verification Latency	Time required for merchants and payment rails to validate proofs affecting transaction processing throughput
Proof Size	Bandwidth requirements and payload impacts on transaction data transmission across payment infrastructures
Privacy Leakage	Formal measurement of information accessible to adversaries through observation of proof artifacts and transaction patterns
Policy Enforcement	Accuracy of spending cap compliance and category control verification across varied constraint configurations
Cross-Rail Interoperability	Validation that identical proof artifacts function across multiple payment infrastructures with minimal adjustment

Table 4: Performance Evaluation Metrics [11]

Conclusion

Delegation protocols incorporating privacy safeguards form critical infrastructure as autonomous commerce expands. Agent authorization frameworks must reconcile accountability requirements with confidentiality needs. This cryptographic approach resolves existing protocol shortcomings by controlling information exposure while preserving compliance checks. The security model ensures transaction validation occurs without revealing spending limits or user details. Implementation across multiple payment systems confirms practical deployment feasibility. Computational analysis shows efficiency levels appropriate for real-world transaction environments. Proof verification completes within acceptable timing thresholds for commercial operations. This work provides standardization guidance supporting industry-wide implementation. Performance data and architectural patterns assist adoption efforts across payment ecosystems. Future enhancements target improved cryptographic circuits handling complex policies, multi-level delegation supporting enterprise scenarios, and integration of adaptive risk evaluation. The design achieves confidential spending controls functioning across varied financial platforms. These cryptographic building blocks enable agent deployment in privacy-sensitive commercial settings. Zero-knowledge methods emerge as a practical infrastructure for commerce requiring strong privacy protections. Balancing transparent verification with hidden constraints resolves inherent conflicts between openness and confidentiality in automated payment systems. Cryptographic delegation becomes foundational infrastructure supporting trustworthy autonomous agents operating under strict privacy requirements in financial contexts.

References

[1] Saurav Bhattacharya, et al., "Enhancing Digital Privacy: The Application of Zero-Knowledge Proofs in Authentication Systems," International Journal of Computer Trends and Technology, April 2024. https://www.researchgate.net/publication/380525014_Enhancing_Digital_Privacy_The_Application_of_Zero-Knowledge_Proofs_in_Authentication_Systems

[2] Sandeep Gupta, "Zero-Knowledge Proofs For Privacy-Preserving Systems: A Survey Across Blockchain, Identity, And Beyond," Engineering and Technology Journal, ResearchGate, July 2025. https://www.researchgate.net/publication/394445573_Zero-Knowledge_Proofs_For_Privacy-Preserving_Systems_A_Survey_Across_Blockchain_Identity_And_Beyond

[3] Sandeep Gupta, "Zero-Knowledge Proofs For Privacy-Preserving Systems: A Survey Across Blockchain, Identity, And Beyond," EVERANT JOURNALS, July 2025. https://everant.org/index.php/etj/article/view/2061

[4] Junliang Liu, Zhiyao Liang, and Qiuyun Lyu, "Empowering Privacy Through Peer-Supervised Self-Sovereign Identity: Integrating Zero-Knowledge Proofs, Blockchain Oversight, and Peer Review Mechanism," MDPI, December 2024. https://www.mdpi.com/1424-8220/24/24/8136

[5] Jothimani Kanthan Ganapathi, "Zero-Knowledge Enabled Cross-Border Payment Systems: Advancing Privacy and Compliance in Blockchain Architectures," Journal of Information Systems Engineering & Management, ResearchGate, August 2025. https://www.researchgate.net/publication/395079777_Zero-Knowledge_Enabled_Cross-Border_Payment_Systems_Advancing_Privacy_and_Compliance_in_Blockchain_Architectures

[6] Geoffrey Goodell, et al., "A Digital Currency Architecture for Privacy and Owner-Custodianship," MDPI, May 2021. https://www.mdpi.com/1999-5903/13/5/130

[7] Jon Watkins, "Zero-Knowledge Proof Techniques for Enhanced Privacy and Scalability in Blockchain Systems," IEEE International Symposium on High-Performance Computer Architecture, ResearchGate, January 2025. https://www.researchgate.net/publication/390034014_Zero-Knowledge_Proof_Techniques_for_Enhanced_Privacy_and_Scalability_in_Blockchain_Systems

[8] Dhruv Patel and Ritesh Tandon, "Cryptographic Trust Models and Zero-Knowledge Proofs for Secure Cloud Access Control and Authentication," International Journal of Advanced Research in Science Communication and Technology, ResearchGate, vol. 2, no. 1, December 2022. https://www.researchgate.net/publication/392795432_Cryptographic_Trust_Models_and_Zero-Knowledge_Proofs_for_Secure_Cloud_Access_Control_and_Authentication

[9] Zhigang Chen, Yuting Jiang, Xinxia Song, and Liqun Chen, "A Survey on Zero-Knowledge Authentication for Internet of Things," MDPI, February 2023. https://www.mdpi.com/2079-9292/12/5/1145

[10] Xin Lin, Yuanyuan Zhang, et al., "An Access Control System Based on Blockchain with Zero-Knowledge Rollups in High-Traffic IoT Environments," National Library of Medicine, March 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC10098902/

[11] Sadaf Mushtaq, et al., "A Systematic Literature Review on the Implementation and Challenges of Zero Trust Architecture Across Domains," MDPI, October 2025.https://www.mdpi.com/1424-8220/25/19/6118