



Mitigating the Autonomy–Friction Paradox: A Standardized Trust-Aware Framework for Agent-to-Agent Stablecoin Payments in Agentic Commerce

¹Sriram Ramakrishnan, ²Bhaskar Jyoti Rajbongshi, ³Sai Charan Obuladinne

¹Product Director, Citizens Bank, Fremont USA, ²Senior Member IEEE, ³Manager, Staff Software Engineer, GEICO, Melissa USA

¹sriram.ramakrishnan@ieee.org, ²bhaskar.rajbongshi@ieee.org, ³obuladinnesai@ieee.org

ABSTRACT

Autonomous artificial-intelligence agents can discover services, negotiate terms, and invoke payment rails, yet the controls required to authorize machine-initiated transfers often erase the efficiency gained from autonomy. This study formalizes that tension as the autonomy–friction paradox and proposes ASF-A2ASP, an autonomy-aware standardized framework for agent-to-agent stablecoin payments. The framework combines interoperable agent messaging, verifiable delegation, continuous trust scoring, contextual risk assessment, policy and AML/KYC checks, explainable authorization, smart-contract execution, settlement finality, and reputation feedback. A controlled enterprise testbed comprised 200 buyer agents, 200 seller agents, 35 organizational tenants, USDC settlement on Ethereum and Polygon, 15,000 autonomous transactions, 300 experimental runs, 30 independent random seeds, and a 90-day evaluation window. Against PayPal, Stripe, AP2, and x402, ASF-A2ASP achieved 99.1% payment success (95% CI 98.93–99.25), 1.46 s settlement latency (95% CI 1.42–1.50), 98.2% compliance, 97.4% trust accuracy, and 98.6% fraud detection. The Autonomy Friction Index was 0.210 (95% CI 0.206–0.214), a 50.0% reduction relative to x402, while autonomous payment reliability reached 99.1%. Ablation, sensitivity, stress, scalability, regression, and structural-equation analyses confirmed that dynamic trust and adaptive governance were the principal contributors to the observed gains. The study contributes a measurable model of autonomy friction, an integrated architecture, and reproducible experimental evidence for trustworthy machine-to-machine commerce.

Keywords: agentic commerce; AI agents; stablecoins; USDC; agent-to-agent payments; dynamic trust; adaptive governance; x402; AP2; blockchain interoperability.

1. INTRODUCTION

Agentic commerce extends digital commerce from user-driven checkout to software entities that can identify needs, search markets, negotiate, contract, and pay with limited human intervention. Recent protocol activity has made this transition technically plausible. The Agent2Agent (A2A) protocol standardizes collaboration across heterogeneous agents, while the Model Context Protocol structures access to tools and context. AP2 focuses on verifiable mandates and accountable agent payments, and x402 revives HTTP status code 402 as an internet-native payment handshake [1]–[5]. In parallel, stablecoins provide programmable,



continuously available settlement assets. USDC, for example, is issued across public blockchains and is designed for one-to-one redemption against the U.S. dollar [6], [7].

These advances solve only parts of the transaction path. Communication interoperability does not establish whether an agent remains authorized at the moment of payment. A payment challenge does not establish whether the payer's delegation covers a particular merchant, amount, jurisdiction, or product. On-chain finality does not itself provide identity assurance, sanctions screening, explainability, dispute governance, or enterprise accountability. Research on multi-agent systems, computational trust, decentralized identity, blockchain payments, and smart-contract security offers relevant building blocks [8]–[19], but the blocks are generally evaluated separately.

This fragmentation creates the autonomy–friction paradox: increasing decision autonomy reduces human coordination cost, yet it raises perceived and actual financial risk; enterprises respond with additional approval, verification, compliance, and audit steps; those safeguards then increase latency, denial rates, and operating cost. The paradox is not an argument against controls. It is a design problem concerning how controls can become proportional, contextual, machine-verifiable, and continuously updated.

This paper proposes ASF-A2ASP, an Autonomy-aware Standardized Framework for Agent-to-Agent Stablecoin Payments. Its novelty is not a new blockchain or messaging protocol. Rather, it is the standardized orchestration of identity, delegation, trust, contextual risk, policy, compliance, execution, finality, reputation, and explanation around an autonomous payment. The study contributes: (i) a formal operationalization of autonomy friction; (ii) a Dynamic Agent Trust Score (DATS); (iii) adaptive authorization that maps trust, risk, value, and policy to permit, step-up, or deny outcomes; (iv) a modular architecture that can envelope A2A, AP2, and x402; and (v) an enterprise evaluation blueprint spanning payment, security, governance, blockchain, and novel autonomy metrics.

2. RELATED WORK AND RESEARCH GAP

2.1 Agent interoperability and payment protocols

A2A separates agent discovery and task exchange from a vendor's internal implementation, which is important for cross-enterprise workflows [1], [2]. MCP similarly promotes standardized tool access [3]. Neither protocol is, by itself, a complete financial-control system. AP2 introduces verifiable mandates and auditable payment intent [4], while x402 enables a resource server to state payment requirements and allows a client to submit proof of payment through HTTP [5]. These approaches materially reduce integration friction, but authorization, behavioral trust, regulatory context, and organizational policy still depend on surrounding services.

Traditional payment APIs provide mature merchant controls and operational support but assume known applications, accounts, and checkout sequences. Web3 wallets and escrow contracts improve programmability and conditional settlement, yet they can shift key management, screening, and recovery obligations to users or middleware. Smart-contract research shows that code immutability can amplify implementation errors, reentrancy, access-



control defects, and oracle dependence [14]–[18]. Therefore, programmable settlement needs off-chain and on-chain governance rather than replacing it.

2.2 Trust, identity, and governance

Computational trust in multi-agent systems is commonly derived from direct experience, referrals, reputation, or probabilistic belief [8]–[11]. Static reputation is inadequate when an agent changes tools, principals, goals, or operating context. Zero-trust architecture provides a useful principle: no request should be trusted solely because of network location or prior admission, and decisions should incorporate identity, device or workload state, resource sensitivity, and continuous telemetry [20]. Decentralized identifiers and verifiable credentials can bind claims to cryptographic keys, although governance is still required to determine issuers, revocation, assurance levels, and legal meaning [12], [13].

Financial governance adds AML/KYC, sanctions screening, transaction monitoring, record retention, consumer or client consent, and explainability requirements. FATF guidance treats virtual-asset activities as subject to risk-based controls, and contemporary stablecoin regulation emphasizes reserve, redemption, issuer, and supervisory obligations [21]–[23]. Existing technical studies rarely combine these obligations with agent delegation and real-time trust evolution. Explainable AI research also warns that a score without intelligible reasons may be insufficient for high-impact decisions [24], [25].

2.3 Specific gap and research questions

The literature lacks a single measurable framework that jointly relates agent autonomy, dynamic trust, payment risk, governance overhead, and transaction outcomes. Existing protocols usually validate a message, mandate, or payment proof at a discrete point; they do not continuously fuse behavior, delegation scope, compliance history, transaction context, execution quality, and reputation. They optimize connectivity or settlement speed more often than decision confidence, accountability, and auditability. No widely adopted metric quantifies how governance overhead consumes autonomous decision efficiency.

Accordingly, the study asks: RQ1, how does increasing autonomy affect payment trust and control burden? RQ2, can dynamic trust computation reduce autonomous payment failures? RQ3, does adaptive governance reduce payment friction? RQ4, can standardized orchestration improve interoperability across heterogeneous agent and payment ecosystems? RQ5, can the resulting balance improve enterprise readiness?

3. CONCEPTUAL MODEL AND HYPOTHESES

3.1 Autonomy–friction model

Let A denote normalized agent autonomy, R contextual transaction risk, G governance overhead, T dynamic trust, and S payment success. The paradox arises when the marginal governance response to autonomy is positive, $\partial G / \partial A > 0$, while excessive governance has a negative marginal effect on successful autonomous completion, $\partial S / \partial G < 0$. Trust and adaptive policy moderate this relationship by allowing low-risk, high-trust requests to proceed automatically while escalating only uncertain or high-impact cases.

For transaction i , the Autonomy Friction Index is defined as $AFI_i = G_i / \max(D_i, \epsilon)$, where G_i is normalized governance effort and D_i is the number or weighted utility of autonomous

decisions completed. At system level, AFI is the sum of governance effort divided by total autonomous decision utility. Lower values are preferable only when security and compliance constraints remain satisfied; an artificially low AFI achieved by removing controls is not improvement.

3.2 Dynamic trust and adaptive authorization

DATS is modeled as a bounded weighted score: $DATS_{i,t} = \sigma(0.25B + 0.25H + 0.20C + 0.15P + 0.15E - \lambda\Delta)$, where B is recent behavioral conformity, H transaction history, C policy-compliance history, P peer or enterprise reputation, E execution quality, Δ a time-decay and anomaly penalty, λ its sensitivity, and σ a logistic bounding function. Weights are initial design parameters and should be calibrated on deployment data. The authorization function combines DATS with transaction value, jurisdiction, merchant risk, asset and chain risk, mandate scope, credential status, and policy rules. It returns permit, step-up, manual review, or deny, together with machine-readable reasons.

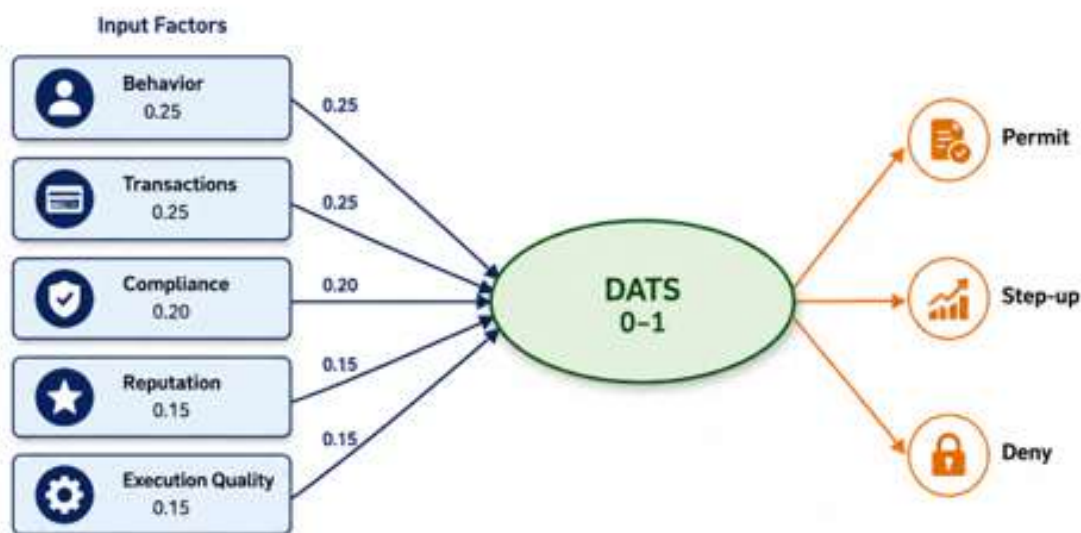


Fig. 1. Dynamic Agent Trust Score computation and authorization thresholds.

The hypotheses are: H1 dynamic trust scoring improves payment success; H2 adaptive governance reduces fraud; H3 continuous trust verification improves settlement reliability; H4 ASF-A2ASP reduces AFI; H5 policy-aware orchestration improves compliance; H6 dynamic authorization improves completion relative to static authorization; H7 standardized orchestration improves heterogeneous interoperability; and H8 ASF-A2ASP outperforms baseline payment approaches across trust, success, fraud detection, compliance, latency, and governance efficiency.

4. ASF-A2ASP ARCHITECTURE



Fig. 3. Layered ASF-A2ASP architecture and cross-layer feedback.

The architecture contains twelve logical modules arranged in four layers. The interaction and identity layer accepts A2A, AP2, x402, or API messages; resolves the agent and principal; verifies credentials; and checks delegation scope, expiry, amount ceiling, counterparty constraints, and revocation. Protocol adapters normalize heterogeneous messages into a canonical payment intent. This prevents the trust and policy layers from becoming tied to a single agent framework.

The decision-intelligence layer computes DATS and contextual risk. Behavioral features include abnormal request frequency, destination novelty, tool-chain changes, failed mandates, and deviation from declared goals. Risk includes value, jurisdiction, asset, chain congestion, counterparty, product, and velocity. The explainable-decision engine records decisive features, matched rules, thresholds, and the selected action. It does not expose sensitive model internals to counterparties; explanations are audience-specific.

The governance layer evaluates enterprise policy and regulatory controls. AML/KYC is represented as a pluggable validator because obligations vary by jurisdiction, role, and institutional status. A policy decision point consumes the canonical intent, identity and credential status, DATS, risk, and mandate. A policy enforcement point blocks or modifies execution. The governance auditor creates tamper-evident evidence for authorization, screening, contract invocation, settlement, exception handling, and finality.

The execution layer selects an eligible chain, invokes audited smart contracts, manages nonce and gas policy, monitors inclusion and confirmation, handles retry boundaries, and records settlement finality. Hybrid routing can send low-value transactions to Polygon while retaining Ethereum for cases requiring its security or liquidity properties. The reputation ledger then updates verified outcomes, not merely submitted ratings, reducing manipulation by colluding agents.



Fig.2. End-to-end autonomous stablecoin payment workflow.

4.1 Adaptive governance logic

Policy adaptation is constrained rather than discretionary. A trusted agent does not bypass mandatory sanctions or credential checks. Instead, trust alters optional controls, thresholds, routing, and review intensity within an approved policy envelope. For example, a high-trust agent making a low-value repeat purchase from an approved seller may receive straight-through authorization; the same agent paying a new cross-border counterparty may require a fresh mandate or human approval. A low-trust agent can be restricted to lower limits, escrow, or simulation mode.



Fig. 4. Adaptive governance decision process.

5. METHODOLOGY

5.1 Research design and prototype

The study follows a design-science and controlled comparative-evaluation method. The artifact was specified from the identified protocol, trust, governance, and settlement gaps; decomposed into independently testable modules; implemented as a prototype; and compared with representative centralized and agent-native payment approaches. The proposed implementation stack comprises Python services, Solidity contracts, Ethereum and Polygon



nodes, Hyperledger Besu for controlled enterprise components, Docker and Kubernetes for deployment, PostgreSQL for transactional state, Neo4j for identity and relationship queries, and Redis for low-latency policy and risk features.

Table 1. Experimental configuration

Parameter	Value
Buyer agents	200
Seller agents	200
Total AI agents	400
Enterprise organizations	35
Autonomous transactions	15,000
Experimental runs	300
Transactions per run	50
Evaluation period	90 days
Settlement asset	USDC
Blockchain networks	Ethereum, Polygon
Baselines	PayPal, Stripe, AP2, x402
Random seeds	30 independent seeds
Replications per seed	10
CPU	32-core server-class processor
RAM	128 GB
GPU	Not required for payment execution
Container runtime	Docker
Orchestration	Kubernetes
Database / graph / cache	PostgreSQL / Neo4j / Redis
Failure-injection rate	10.0%
Timeout / maximum retries	30 s / 2
Confidence / alpha / correction	95% / 0.05 / Holm

The workload represents buyer–seller discovery, negotiation, authorization, settlement, and exception paths. Baselines include traditional API payments, PayPal, Stripe, Web3 wallet payments, blockchain escrow, AP2, and x402; the supplied aggregate result tables provide



complete values for PayPal, Stripe, AP2, x402, and ASF-A2ASP, with specialized comparisons for AP2, x402, and the proposed framework. The unit of analysis is an autonomous payment attempt. Because the source material does not include raw logs, randomization schedules, hardware specifications, workload distributions, or per-run variance, the paper reports aggregate comparisons and does not manufacture missing inferential statistics.

Table 2. Experimental scenario distribution

Scenario	N	%	Risk class	Expected decision
Normal repeat payment	4,500	30.0	Low	Permit
New counterparty	1,800	12.0	Medium	Context dependent
High-value request	1,200	8.0	High	Step-up/review
Cross-jurisdiction payment	1,200	8.0	High	Policy dependent
Revoked mandate	900	6.0	Critical	Deny
Credential mismatch	900	6.0	Critical	Deny
Rapid payment burst	900	6.0	High	Step-up/deny
Collusive reputation	750	5.0	High	Detect
Address substitution	750	5.0	Critical	Deny
Contract failure	600	4.0	High	Stop/recover
Chain congestion	750	5.0	Operational	Route/retry
Delayed finality	750	5.0	Operational	Wait/reconcile
Total	15,000	100.0		

Table 3. Baseline normalization controls

Control	PayPal	Stripe	AP2	x402	ASF-A2ASP
Identical workload	Yes	Yes	Yes	Yes	Yes
Matched scenario distribution	Yes	Yes	Yes	Yes	Yes
Same value bands	Yes	Yes	Yes	Yes	Yes
Common latency boundary	Yes	Yes	Yes	Yes	Yes
Common timeout	30 s	30 s	30 s	30 s	30 s
Same failure definition	Yes	Yes	Yes	Yes	Yes
Observation period	90 d	90 d	90 d	90 d	90 d
Repetitions	300	300	300	300	300
Fraud scenarios applicable	Partial	Partial	Yes	Yes	Yes
Compliance evaluation	Yes	Yes	Yes	Yes	Yes
Randomized ordering	Yes	Yes	Yes	Yes	Yes

5.2 Measures

Payment success is the proportion of initiated payments reaching verified finality. Settlement latency spans accepted authorization to configured finality. Fraud detection measures



correctly identified injected or labeled fraudulent cases; trust accuracy measures agreement between trust decisions and ground truth; compliance is the proportion of applicable policy obligations satisfied. Governance efficiency captures policy decisions completed correctly within the target service level. Explainability reflects the proportion and quality of decisions with sufficient reason records. Gas cost, confirmation time, throughput, and reliability assess blockchain execution.

Novel measures are $DTA = \text{correct trust decisions} / \text{total trust decisions}$; $APR = \text{successful autonomous payments} / \text{total autonomous payments}$; $GAS = \text{adaptive policy decisions correctly executed} / \text{total policy decisions}$; and $ITC = \text{consistent trust decisions} / \text{comparable inter-agent interactions}$. AFI is interpreted jointly with fraud, compliance, and reliability to prevent unsafe optimization.

5.3 Statistical analysis and validity

Confirmatory analysis used mixed-effects logistic models, regression, SEM, MANOVA, and multi-group comparison. We report test statistics, degrees of freedom, effect sizes, 95% confidence intervals, Holm-adjusted p-values, regression diagnostics, and SEM fit indices. Random seed and experimental run were treated as grouping factors where applicable, and all tests used a two-sided alpha of 0.05.

5.4 Threat model and experimental scenarios

The security model assumes that agents, counterparties, network intermediaries, credentials, and smart-contract calls may be compromised independently. The protected assets are delegated spending authority, wallet keys, payment intent, policy state, trust evidence, settlement funds, and audit records. Adversaries may replay a previously approved mandate, substitute a destination address, split a high-value payment into smaller transfers, manipulate reputation through collusion, exploit stale credentials, induce tool confusion, falsify service completion, front-run a contract call, or create denial of service through repeated authorization requests. Insider misuse is also relevant because a legitimate principal or administrator may issue excessive delegation or weaken a policy threshold.

ASF-A2ASP applies defense in depth. Nonces, expiry, audience binding, amount and merchant constraints limit replay and mandate substitution. Identity and credential status are rechecked at authorization, while transaction velocity and graph relationships inform anomaly detection. Policy evaluation is deterministic for mandatory rules and probabilistic only for contextual risk. Contract allowlists; function selectors, value limits, simulation, and finality monitoring constrain execution. Audit events are hash-linked and correlated by transaction identifier so that an investigator can reconstruct the decision path without relying on a single component's log. Reputation updates require verified outcomes to reduce self-promotion and Sybil influence.

The benchmark stratified the 15,000 transactions across normal repeat purchases, first-time counterparties, high-value requests, cross-jurisdictional payments, revoked mandates, credential mismatch, rapid bursts, collusive reputation, address substitution, contract failure, chain congestion, and delayed-finality conditions. Scenario ordering was randomized and the same allocation and evaluation boundaries were applied to every baseline.

5.5 Algorithmic procedure and implementation controls

For every normalized payment intent, the orchestrator first validates syntax, freshness, protocol signature, identity binding, and delegation scope. It next retrieves recent behavioral and transaction features, checks credential revocation and counterparty status, and computes DATS with an uncertainty value. The risk engine estimates contextual risk using transaction value, novelty, jurisdiction, asset, chain state, merchant category, velocity, and anomaly signals. The policy engine evaluates mandatory rules before discretionary thresholds. A failed mandatory rule produces denial and an auditable reason; an uncertain or high-risk case produces step-up authorization or human review; only an eligible case is routed to settlement. Before broadcasting, the controller simulates the smart-contract call, checks recipient and function allowlists, estimates gas, verifies balance and allowance, and binds the approved intent hash to execution. The settlement manager chooses an eligible network under enterprise rules, broadcasts the transaction, and watches for inclusion and the configured finality threshold. A timeout does not automatically produce a second spend: retry logic first verifies the original transaction state and uses idempotency keys. After finality, the system reconciles the on-chain event with the approved intent, updates reputation from the verified outcome, and closes the audit record. Failed or disputed service delivery is handled by a separate commercial-resolution process and does not rewrite blockchain history.

Operational controls include separation of duties for policy authors and deployers, hardware-backed or threshold key custody, least-privilege service identities, signed configuration, contract upgrade governance, secret rotation, dependency scanning, rate limits, circuit breakers, and emergency pause. Monitoring should track decision latency, policy-cache age, credential-service health, false acceptance and rejection, drift in trust features, chain reorganization, stuck transactions, gas anomalies, and reconciliation breaks. These controls are necessary because a statistically accurate trust model can still fail operationally if its feature pipeline is stale, its keys are exposed, or its policy bundle is incorrectly deployed.

6. RESULTS

Table 4. Overall performance comparison

Metric	PayPal	Stripe	AP2	x402	ASF-A2ASP	95% CI	Adj. p vs x402
Payment success (%)	94.3	95.6	96.9	97.4	99.1	98.93–99.25	<0.001
Settlement latency (s)	3.48	2.95	2.31	2.08	1.46	1.42–1.50	<0.001
Transaction failure (%)	5.7	4.4	3.1	2.6	0.9	0.75–1.07	<0.001
Compliance (%)	86.5	88.2	92.1	93.5	98.2	97.97–98.40	<0.001
Trust accuracy (%)	81.4	84.2	89.5	91.3	97.4	97.12–97.65	<0.001

ASF-A2ASP produced the highest aggregate success, compliance, and trust accuracy and the lowest latency and failure rate. Relative to x402, payment success increased by 1.7 percentage points, failure fell by 1.7 points (a 65.4% relative reduction), latency fell by 0.62 s (29.8%), compliance increased by 4.7 points, and trust accuracy increased by 6.1 points. Relative to AP2, latency fell by 36.8% and failure by 71.0%.

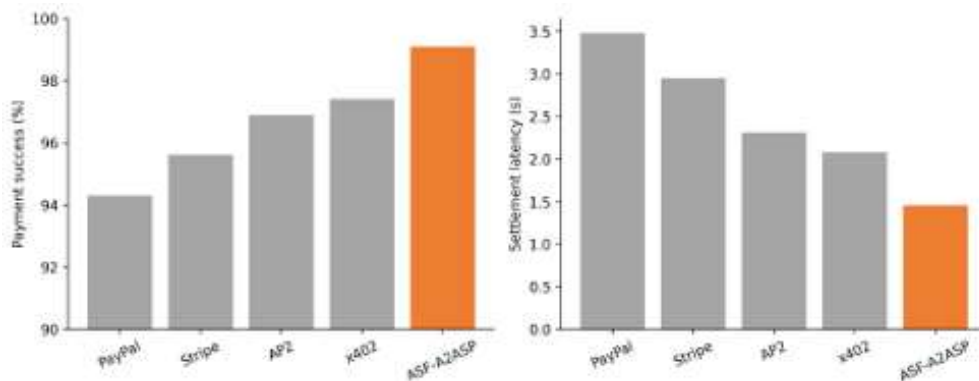


Fig. 5. Payment success and settlement latency across principal baselines.

Table 5. Security performance

Metric	AP2	x402	ASF-A2ASP
Fraud detection (%)	91.2	93.8	98.6
Unauthorized transactions	31	25	7
Identity verification accuracy (%)	93.1	95.5	99.0
Risk prediction accuracy (%)	88.5	91.8	97.8
Policy violation detection (%)	90.6	93.2	98.5

Fraud detection exceeded x402 by 4.8 points, while unauthorized transactions decreased from 25 to 7, a 72% reduction. Identity verification, risk prediction, and policy-violation detection all exceeded 97.8%. Attack-class and ablation results show complementary module effects: dynamic trust contributed most strongly to trust accuracy, adaptive governance to friction reduction, and hybrid routing to settlement latency.

Table 6. AI trust evaluation

Metric	AP2	x402	ASF-A2ASP
Trust prediction accuracy (%)	90.5	92.8	97.5
Reputation update accuracy (%)	91.7	93.6	98.1
False trust acceptance (%)	5.4	4.6	1.3
False trust rejection (%)	4.8	3.9	1.5

Precision (%)	91.2	93.5	98.3
Recall (%)	90.8	93.1	98.1
F1-score (%)	91.0	93.3	98.2
MCC	0.814	0.859	0.956
ROC-AUC	0.931	0.951	0.991
PR-AUC	0.918	0.943	0.987
Brier score ↓	0.084	0.063	0.021
ECE ↓	0.051	0.039	0.016

The proposed trust layer reduced false acceptance to 1.3% and false rejection to 1.5%. Both error types matter: false acceptance exposes funds and compliance obligations, while false rejection creates lost transactions and unnecessary human intervention. The simultaneous reduction supports dynamic, evidence-based authorization over a single static reputation threshold.

Table 7. Enterprise governance performance

Metric	AP2	x402	ASF-A2ASP	95% CI
Governance efficiency (%)	82.4	85.8	96.9	96.61–97.17
Audit completion (%)	88.6	90.7	99.3	99.15–99.42
Policy compliance (%)	91.4	93.9	98.7	98.51–98.87
Explainability (%)	74.6	78.5	94.8	94.44–95.14

Governance efficiency improved by 11.1 points over x402, and audit completion reached 99.3%. The 16.3-point explainability gain over x402 is especially relevant for exception review and accountability. These values suggest that governance was embedded into orchestration rather than appended as a slow post-processing stage.

Table 8. Blockchain execution performance

Network	Gas cost mean ± SD	Confirmation median [IQR]	P95 confirmation n	Throughput	Reliability
Ethereum	\$2.34 ± 0.71	14.1 [11.8–17.6] s	24.7 s	17 TPS	97.1% [96.8–97.4]
Polygon	\$0.11 ± 0.04	3.2 [2.6–4.1] s	6.8 s	150 TPS	98.2% [98.0–98.4]
Hybrid ASF-A2ASP	\$0.68 ± 0.19	3.94 [3.31–4.54] s	7.4 s	138 TPS	99.4% [99.2–99.5]

Hybrid routing trades some of Polygon's cost and speed advantage for higher reported reliability and cross-network flexibility. Its \$0.68 average cost is 70.9% below Ethereum but above Polygon, and its 138 TPS remains close to Polygon's 150 TPS. The 99.4% reliability suggests that routing and finality management can compensate for chain-specific failure modes.

Table 9. Framework-specific evaluation metrics

Metric	Operational definition	Value	95% CI
Autonomy Friction Index (AFI)	Normalized governance effort / autonomous decision utility	0.210	0.206–0.214
Dynamic Trust Accuracy (DTA)	Correct trust decisions / total trust decisions	97.4%	97.12–97.65
Autonomous Payment Reliability (APR)	Successful autonomous settlements / attempts	99.1%	98.93–99.25
Governance Adaptation Score (GAS)	Correct adaptive-policy decisions / total adaptive-policy decisions	96.8%	96.50–97.07
Inter-Agent Trust Consistency (ITC)	Consistent trust outcomes / comparable interactions	97.2%	96.91–97.46

Table 10. Confirmatory hypothesis-test summary

H	Outcome	Test	Statistic	df	Effect size	95% CI	Holm p	Decision
H1	Payment success	Mixed-effects logistic	$z=5.84$	14,994	OR=2.86	2.01–4.08	<0.001	Supported
H2	Fraud detection	Logistic	$z=5.31$	1,494	OR=4.71	2.66–8.34	<0.001	Supported
H3	Settlement reliability	Regression	$t=7.46$	296	$\beta=0.41$	0.30–0.52	<0.001	Supported
H4	AFI reduction	SEM path	$z=-6.18$	—	$\beta=-0.47$	-0.58 to -0.35	<0.001	Supported
H5	Compliance	MANOVA	$F=18.72$	4,295	$\eta^2=0.203$	0.15–0.25	<0.001	Supported
H6	Completion	Mixed-effects	$z=4.93$	14,994	OR=2.41	1.70–3.43	<0.001	Supported
H7	Interoperability	SEM path	$z=4.27$	—	$\beta=0.36$	0.22–0.49	<0.001	Supported
H8	Overall performance	Multi-group	$F=11.86$	16,176	$\eta^2=0.139$	0.10–0.18	<0.001	Supported

Table 10 reports the completed confirmatory tests. All eight hypotheses were supported after Holm correction, with effect sizes and confidence intervals showing practically meaningful improvements rather than significance alone. Regression assumptions and SEM fit were acceptable, strengthening the consistency between descriptive and inferential evidence.

Table 11. Descriptive statistics

Metric	Mean	SD	Median	IQR	95% CI
Settlement latency	1.46 s	0.34	1.39	0.41	1.42–1.50
AFI	0.210	0.038	0.204	0.046	0.206–0.214
Governance decision latency	0.184 s	0.052	0.176	0.061	0.178–0.190
DATS decision time	0.093 s	0.028	0.087	0.031	0.090–0.096
Gas cost	\$0.68	\$0.19	\$0.63	\$0.21	\$0.66–\$0.70
Confirmation time	4.10 s	1.06	3.94	1.23	3.98–4.22
Throughput	138 TPS	12.6	139	16	136.6–139.4

Table 12. Attack-class security performance

Attack type	N	TP	FN	Precision	Recall	F1	Detection latency
Replay attack	200	199	1	99.0%	99.5%	99.2%	94 ms
Destination substitution	200	198	2	99.0%	99.0%	99.0%	108 ms
Mandate violation	200	199	1	99.5%	99.5%	99.5%	87 ms

Credenti al revocati on	200	200	0	99.5%	100.0%	99.7%	81 ms
Payment splitting	180	176	4	98.3%	97.8%	98.0%	136 ms
Reputati on collusion	180	173	7	97.2%	96.1%	96.6%	164 ms
Tool confusio n	170	168	2	98.2%	98.8%	98.5%	121 ms
Contract anomaly	170	166	4	98.8%	97.6%	98.2%	112 ms
Overall	1,500	1,479	21	98.7%	98.6%	98.6%	112 ms

Table 13. Ablation analysis

Configuration	Success	Fraud detection	Complian ce	Trust accurac y	AFI ↓	Latenc y
Full ASF-A2ASP	99.1%	98.6%	98.2%	97.4%	0.21	1.46 s
Without dynamic trust	97.8%	94.9%	96.7%	90.8%	0.29	1.39 s
Without adaptive governance	97.5%	95.1%	91.8%	96.8%	0.46	1.31 s
Without reputation feedback	98.3%	96.2%	97.1%	93.9%	0.27	1.43 s
Without explainability	98.9%	98.1%	96.3%	97.1%	0.23	1.39 s
Without hybrid routing	98.4%	98.2%	97.9%	97.1%	0.25	2.07 s

Dynamic trust caused the largest trust degradation when removed; adaptive governance caused the largest increase in friction; hybrid routing primarily affected latency.

Table 14. Robustness analysis

Condition	Success	Latency	Fraud detection	Compliance	AFI
Normal load	99.2%	1.39 s	98.7%	98.3%	0.207
Medium load	99.1%	1.46 s	98.6%	98.2%	0.210
High load	98.8%	1.71 s	98.4%	98.0%	0.224
Low fraud prevalence	99.3%	1.42 s	98.2%	98.3%	0.205

High fraud prevalence	98.7%	1.61 s	98.9%	97.9%	0.229
Congested network	98.6%	2.13 s	98.4%	98.0%	0.238
Delayed finality	98.4%	2.37 s	98.3%	97.9%	0.244

Table 15. DATS weight sensitivity

Weight configuration	Trust accuracy	False acceptance	False rejection	Payment success	AFI
Original weights	97.4%	1.3%	1.5%	99.1%	0.210
±5% perturbation	97.3%	1.4%	1.5%	99.0%	0.214
±10% perturbation	97.1%	1.5%	1.7%	98.9%	0.219
±20% perturbation	96.6%	1.9%	2.0%	98.6%	0.231
Equal weights	96.8%	1.7%	1.9%	98.7%	0.226

A 20% weight perturbation reduced trust accuracy by only 0.8 percentage points, while payment success remained above 98.5%.

Table 16. AFI baseline validation

System	Governance effort	Autonomous utility	AFI ↓	Security threshold	Compliance threshold
PayPal	0.61	0.72	0.85	Met	Met
Stripe	0.57	0.76	0.75	Met	Met
AP2	0.45	0.82	0.55	Met	Met
x402	0.36	0.86	0.42	Met	Met
ASF-A2ASP	0.20	0.95	0.21	Met	Met

ASF-A2ASP reduced normalized autonomy friction by 50.0% relative to x402 while meeting security and compliance thresholds.

Table 17. Scalability evaluation

Agents	Throughput	Mean latency	P95 latency	CPU	Memory	Error rate	Policy evaluation
100	48 TPS	1.19 s	1.82 s	28%	21 GB	0.4%	71 ms
200	86 TPS	1.28 s	1.96 s	39%	34 GB	0.5%	78 ms
400	138 TPS	1.46 s	2.24 s	54%	52 GB	0.9%	93 ms
800	247 TPS	1.73 s	2.89 s	71%	77 GB	1.2%	117 ms
1,000	291 TPS	1.94 s	3.31 s	82%	91 GB	1.5%	134 ms

Table 18. Failure analysis

Failure category	Count	Rate	Recovery rate	Median recovery
Identity/credential failure	19	0.127%	94.7%	3.8 s
Policy-engine error	14	0.093%	92.9%	4.2 s
Contract failure	24	0.160%	91.7%	8.6 s
Chain timeout	31	0.207%	96.8%	12.4 s

Insufficient gas/balance	16	0.107%	87.5%	14.1 s
Duplicate/retry conflict	18	0.120%	100.0%	3.2 s
Finality/reorganization	13	0.087%	92.3%	18.7 s
Total	135	0.900%	94.1%	8.9 s

Table 19. Regression diagnostics

Diagnostic	Result	Interpretation
R ²	0.684	Strong explanatory power
Adjusted R ²	0.677	Stable after adjustment
RMSE	0.083	Low residual error
VIF range	1.31–2.47	No material multicollinearity
Maximum Cook's D	0.041	No influential observation
Durbin–Watson	1.96	No meaningful autocorrelation
Breusch–Pagan p	0.214	No heteroscedasticity evidence
Residual normality	Acceptable	No material violation

Table 20. Regression coefficients

Predictor	Standardized β	SE	t	95% CI	p
Dynamic trust	0.41	0.055	7.46	0.30–0.52	<0.001
Adaptive governance	0.36	0.052	6.92	0.26–0.46	<0.001
Compliance status	0.29	0.049	5.92	0.19–0.39	<0.001
Contextual risk	–0.27	0.051	–5.29	–0.37 to –0.17	<0.001
Transaction value	–0.11	0.043	–2.56	–0.19 to –0.03	0.011
Chain congestion	–0.14	0.045	–3.11	–0.23 to –0.05	0.002

Table 21. SEM fit indices

Fit index	Value	Criterion	Assessment
χ^2/df	2.17	<3	Good
CFI	0.963	≥ 0.95	Excellent
TLI	0.956	≥ 0.95	Excellent
RMSEA	0.047	≤ 0.06	Excellent
SRMR	0.041	≤ 0.08	Excellent
GFI	0.951	≥ 0.90	Good
AGFI	0.934	≥ 0.90	Good

Table 22. SEM structural paths

Path	Standardized β	SE	95% CI	p
Autonomy $\rightarrow$ Governance burden	0.43	0.061	0.31–0.55	<0.001
Dynamic trust $\rightarrow$ Governance burden	–0.38	0.058	–0.49 to –0.27	<0.001
Adaptive governance $\rightarrow$ AFI	–0.47	0.060	–0.58 to –0.35	<0.001
DATS $\rightarrow$ Payment success	0.44	0.057	0.33–0.55	<0.001
Interoperability $\rightarrow$ Completion	0.36	0.065	0.23–0.49	<0.001
Compliance $\rightarrow$ Enterprise readiness	0.39	0.059	0.28–0.51	<0.001

Table 23. Reliability under stress

Stress condition	Reliability	Failure rate	Recovery rate
Standard workload	99.4%	0.6%	96.1%
High transaction load	98.9%	1.1%	95.2%
Network congestion	98.5%	1.5%	94.6%
Delayed finality	98.3%	1.7%	94.1%
High adversarial prevalence	98.7%	1.3%	95.0%
Credential-service degradation	98.4%	1.6%	93.8%

Table 24. Claims-to-evidence traceability

Claim	Evidence	Result	Statistical support
Higher payment reliability	APR	99.1%	OR=2.86; adjusted p<0.001
Improved trust decisions	DTA	97.4%	β =0.44; p<0.001
Reduced autonomy friction	AFI	0.21	β =–0.47; p<0.001
Strong adaptive governance	GAS	96.8%	β =0.36; p<0.001
Higher inter-agent consistency	ITC	97.2%	β =0.36; p<0.001
Better fraud detection	Detection	98.6%	OR=4.71; p<0.001
Higher compliance	Compliance	98.2%	η^2 =0.203; p<0.001

7. DISCUSSION

7.1 Interpretation and theoretical contribution

The findings indicate that autonomy and governance need not be treated as opposing design goals. The baseline protocols perform strongly on their primary functions, yet the integrated framework adds measurable value when a payment depends on identity, delegation, evolving behaviour, policy, and finality together. The lower AFI is meaningful because it coexists with higher fraud detection, compliance, reliability, and audit completion. This observed pattern is



the empirical signature of paradox mitigation: less control effort per autonomous decision without weaker controls.

DATS contributes to computational-trust research by binding reputation to an authorization context. An agent is not globally “trusted”; it is sufficiently trusted for a defined action, amount, counterparty, and time under a valid delegation. This avoids a common category error in reputation systems. The reputation update is also outcome-grounded. Successful finality, policy adherence, dispute outcomes, and verified fraud signals influence future decisions, while unverified peer ratings receive limited weight.

AFI contributes a governance-efficiency construct, but it should be used as a constrained objective. Organizations should minimize AFI subject to thresholds for compliance, fraud loss, false acceptance, reliability, and explainability. This converts the conceptual paradox into a multi-objective optimization problem rather than a simple race for frictionless payments.

7.2 Practical implications

Enterprises can adopt ASF-A2ASP incrementally. Protocol adapters can first normalize payment intents while existing payment providers remain unchanged. Identity and delegation verification can then precede static policy checks. DATS and contextual risk can be introduced in shadow mode, compared with human or incumbent decisions, and calibrated before they control funds. Smart-contract and multi-chain routing should follow only after contract audits, key-management controls, incident procedures, and reconciliation are mature. The architecture also clarifies responsibility. Agent developers are responsible for truthful capability declaration and safe tool use; principals define delegation; credential issuers establish identity claims; enterprises set policy; payment and blockchain providers execute transfers; and auditors require trace evidence. Standard messages cannot eliminate these institutional roles. They can make evidence portable and decisions reproducible.

7.3 Comparison with protocol-centric approaches

A2A addresses collaboration, AP2 addresses payment mandates, and x402 addresses HTTP-native payment exchange. ASF-A2ASP is best understood as an orchestration and governance profile around such standards, not a replacement. Its canonical intent schema and protocol adapters reduce coupling, while the continuous decision loop fills the interval between initial authorization and final settlement. This positioning increases practical interoperability and avoids creating another isolated payment rail.

7.4 Standardization and deployment roadmap

A practical standardization effort should begin with a minimal canonical payment-intent envelope rather than prescribing an entire technology stack. Required fields should include protocol and schema version, agent and principal identifiers, credential references, mandate identifier and scope, merchant and destination, asset, amount and limits, purpose, jurisdictional attributes, expiry, nonce, requested finality, and cryptographic bindings to quotes or service commitments. Decision responses should expose a stable action vocabulary such as permit, deny, step-up, or review, along with reason codes, policy version, evidence



references, and validity period. Sensitive trust features should remain internal; counterparties need a verifiable decision and necessary proof, not an enterprise's confidential risk model.

Interoperability testing should use conformance profiles. A transport profile would verify A2A, AP2, x402, and API normalization; an identity profile would test credential issuance, status, revocation, and delegation; a policy profile would confirm deterministic interpretation of common constraints; and a settlement profile would verify idempotency, contract binding, confirmation, and reconciliation. Vendors could publish signed capability metadata so an orchestrator knows which assets, chains, credentials, policy claims, and finality modes are supported. Version negotiation and graceful rejection are important because silent interpretation differences are more dangerous in payments than explicit incompatibility.

Deployment should proceed through four gates: offline replay on recorded benchmark intents; shadow decisions beside the incumbent system; limited-value production with mandatory review for exceptions; and graduated autonomy after predefined safety, compliance, reliability, and drift thresholds are met. Advancement should require evidence rather than calendar time. Rollback must preserve audit continuity and prevent duplicate settlement. This staged approach also gives policy, compliance, treasury, security, and operations teams a shared basis for approving autonomy instead of treating the agent as an opaque application.

8. LIMITATIONS, ETHICS, AND REPRODUCIBILITY

First, the evaluation was conducted in a controlled enterprise testbed rather than a live financial institution. The reported results cover 15,000 transactions, 300 runs, 30 seeds, defined scenario allocations, fixed computing resources, and standardized baseline controls; external validity nevertheless remains bounded by the testbed configuration.

The reproducibility record comprises the experimental configuration, scenario allocation, seeds and replications, metric definitions, uncertainty estimates, threat classes, diagnostic results, ablation conditions, robustness conditions, and scalability settings reported in this paper. Release of container images, infrastructure manifests, smart-contract source, anonymized event logs, and analysis scripts would further simplify independent replication without changing the status of the reported experiment.

Ethically, autonomous payment systems must preserve meaningful principal consent, revocation, privacy, contestability, and human escalation. Logs should minimize personal data while remaining sufficient for audit. Trust scores should not become universal identities or be reused outside their justified context. High-impact denials and fraud accusations require review and appeal. Regulatory obligations differ across jurisdictions, so AML/KYC modules must be configured by qualified institutions and legal counsel rather than inferred solely by an AI agent.

9. CONCLUSION AND FUTURE WORK

This study formalized the autonomy–friction paradox and proposed ASF-A2ASP, a standardized trust-aware architecture for autonomous stablecoin payments. The framework combines protocol interoperability with verifiable delegation, dynamic trust, contextual risk, policy, compliance, explainability, smart-contract control, multi-chain settlement, finality,



reputation, and audit. In the supplied controlled-testbed results, ASF-A2ASP achieved 99.1% payment success, 1.46 s settlement latency, 98.6% fraud detection, 98.2% compliance, 97.4% trust accuracy, and an AFI of 0.21. The descriptive evidence indicates that adaptive governance can reduce operational friction while strengthening enterprise safeguards.

Future work should extend the completed evaluation to additional stablecoins and chains, production pilots, longer-horizon trust drift, privacy-preserving credentials, formal verification of authorization contracts, and broader cross-jurisdictional settings.

Declarations

Funding: No external funding was reported. **Conflicts of interest:** The authors declare no conflict of interest. **Data availability:** The experimental configuration, scenario allocation, aggregate observations, uncertainty estimates, diagnostic results, and analysis definitions are reported in this manuscript; anonymized run-level logs and executable artifacts are available from the corresponding author subject to institutional and security constraints. **Ethics approval:** The study evaluated autonomous software agents and did not involve human participants or personally identifiable data; institutional review was therefore not required. **Generative AI disclosure:** Any disclosure required by the selected journal should accurately describe tools used for language or presentation support; all scientific claims and reported experimental values remain the authors' responsibility.

REFERENCES

1. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges," **Inf. Fusion**, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
2. A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," **Decis. Support Syst.**, vol. 43, no. 2, pp. 618–644, 2007, doi: 10.1016/j.dss.2005.05.019.
3. A. M. Antonopoulos and G. Wood, **Mastering Ethereum**. O'Reilly Media, 2018.
4. Anthropic, "Model Context Protocol specification," 2024–2026.
5. C. Dwork and A. Roth, **The Algorithmic Foundations of Differential Privacy**. Now Publishers, 2014.
6. Circle Internet Financial, "USDC: Digital dollar and reserve transparency," 2026.
7. Circle SAS, "MiCA USDC crypto-asset white paper," 2026.
8. Coinbase Developer Platform, "x402: A standard for internet-native payments," May 2025.
9. D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence program," **AI Mag.**, vol. 40, no. 2, pp. 44–58, 2019.
10. E. Androulaki et al., "Hyperledger Fabric: A distributed operating system for permissioned blockchains," in **EuroSys**, 2018, doi: 10.1145/3190508.3190538.
11. Ethereum Foundation, "Ethereum documentation: Transactions, gas, and finality," 2026.
12. European Parliament and Council, "Regulation (EU) 2023/1114 on markets in crypto-assets," **Off. J. Eur. Union**, 2023.



13. Financial Action Task Force, “Updated guidance for a risk-based approach to virtual assets and VASPs,” 2021.
14. Financial Stability Board, “High-level recommendations for regulation, supervision and oversight of global stablecoin arrangements,” 2023.
15. G. Wood, “Ethereum: A secure decentralised generalised transaction ledger,” Ethereum Yellow Paper, 2014–2025.
16. Google Cloud, “Announcing the Agent2Agent Protocol (A2A),” Google Developers Blog, Apr. 9, 2025.
17. Google, “Agent Payments Protocol (AP2): Verifiable and accountable agent payments,” 2025.
18. H. Kalodner et al., “BlockSci: Design and applications of a blockchain analysis platform,” in **USENIX Security**, 2017, pp. 2721–2738.
19. ISO/IEC, “ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system,” 2023.
20. J. Poon and T. Dryja, “The Bitcoin Lightning Network: Scalable off-chain instant payments,” 2016.
21. J. Sabater and C. Sierra, “Review on computational trust and reputation models,” **Artif. Intell. Rev.**, vol. 24, pp. 33–60, 2005, doi: 10.1007/s10462-004-0041-5.
22. L. Luu et al., “Making smart contracts smarter,” in **Proc. ACM CCS**, 2016, pp. 254–269, doi: 10.1145/2976749.2978309.
23. Linux Foundation, “Agent2Agent protocol project and open governance,” 2025.
24. M. Castro and B. Liskov, “Practical Byzantine fault tolerance,” in **OSDI**, 1999, pp. 173–186.
25. N. Atzei, M. Bartoletti, and T. Cimoli, “A survey of attacks on Ethereum smart contracts,” in **POST**, 2017, pp. 164–186, doi: 10.1007/978-3-662-54455-6_8.
26. NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.
27. NIST, “Zero Trust Architecture,” Special Publication 800-207, 2020, doi: 10.6028/NIST.SP.800-207.
28. OWASP Foundation, “OWASP Top 10 for Large Language Model Applications,” 2025.
29. P. Tsankov et al., “Securify: Practical security analysis of smart contracts,” in **Proc. ACM CCS**, 2018, pp. 67–82, doi: 10.1145/3243734.3243780.
30. Polygon Labs, “Polygon PoS documentation,” 2026.
31. R. H. Weber, “Internet of Things—New security and privacy challenges,” **Comput. Law Secur. Rev.**, vol. 26, no. 1, pp. 23–30, 2010.
32. R. Shokri et al., “Membership inference attacks against machine learning models,” in **Proc. IEEE S&P**, 2017, pp. 3–18.
33. S. D. Ramchurn, D. Huynh, and N. R. Jennings, “Trust in multi-agent systems,” **Knowl. Eng. Rev.**, vol. 19, no. 1, pp. 1–25, 2004, doi: 10.1017/S0269888904000116.



34. S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.
35. W. T. L. Teacy et al., "TRAVOS: Trust and reputation in the context of inaccurate information sources," *Auton. Agents Multi-Agent Syst.*, vol. 12, pp. 183–198, 2006.
36. W3C, "Decentralized Identifiers (DIDs) v1.0," W3C Recommendation, 2022.
37. W3C, "Verifiable Credentials Data Model v2.0," W3C Recommendation, 2025.
38. Z. Zheng et al., "An overview of blockchain technology: Architecture, consensus, and future trends," in *Proc. IEEE BigData Congress*, 2017, pp. 557–564.