https://doi.org/10.2352/EI.2026.38.3.MOBMU-331

Cross-Protocol Attack Surface Analysis in Multi-Agent AI Systems: Formal Verification and Categorical Modelling of Translation-Boundary Vulnerabilities

Mahipal

Casky.AI Inc., 2000 Oakley Park Road, Ste 230-C, Commerce Township, MI 48390, United States of America
Email: mahipal@casky.ai

Keywords: Multi-agent AI security, cross-protocol translation, formal verification, category theory, CPRV, MCP, A2A, FIPA-ACL, attack surface analysis, ProVerif, Tamarin, TLA+

Abstract

Three incompatible communication protocols now dominate multi-agent AI deployments: the Model Context Protocol (MCP), Google's Agent-to-Agent Protocol (A2A), and FIPA-ACL. Yet every production system that bridges two of them introduces a translation boundary whose security properties no existing framework examines. This paper presents CPASA, a cross-protocol attack surface analysis with five interlocking results. A category-theoretically grounded taxonomy defines six vulnerability categories, foremost among them the novel Cross-Protocol Replay Vulnerability *(CPRV), and models the six directed translation paths as functors between protocol categories, with security properties expressed as natural transformations.* ***87 machine-checkable formal verdicts*** *(44 ProVerif, 36 Tamarin, 7 TLA+) yield 77 verified properties and 10 violations; three independent formalisms—category theory, ProVerif, and TLA+—with corroborating Tamarin evidence converge on a single conclusion: the MCP–FIPA–MCP roundtrip irreversibly strips OAuth 2.1 credentials, constituting CPRV. A 120-feature ML classifier (macro-F1 = 0.72) provides complementary real-time detection when data are available. Systematic enumeration of 80 attack vectors confirms CPRV as a vulnerability class absent from MAESTRO, OWASP, AgentDojo, and all prior threat modelling. We believe the central finding here is architectural in nature. FIPA-ACL was designed in 2002 without cryptographic authentication, and no amount of gateway engineering can restore credentials that a translation functor has already destroyed—this is not a gap that better statistics will close. The formal proofs do not expire. The ML metrics, on the other hand, depend heavily on the data distribution available at training time and should be treated accordingly.*

Introduction

Protocol translation kills authentication. That single sentence summarises the most consequential finding of this research, and it took 87 formal verification runs across three independent tools to establish it with machine-checkable certainty. The statement sounds hyperbolic until one traces what actually happens when an OAuth 2.1-authenticated MCP request crosses into the FIPA-ACL domain: the token is silently discarded—FIPA has no field for it—and the return translation synthesises a `ProcessCredential` with `verified=false`. The original credential is gone. Not degraded, not downgraded: destroyed.

The agentic AI landscape in early 2026 is split across three protocol families that were never designed to interoperate. Anthropic's Model Context Protocol has surpassed 97 million monthly SDK downloads since its donation to the Linux Foundation in December 2025 [1]. Google's Agent-to-Agent Protocol has secured adoption from over 150 enterprises, absorbing the Agentive Communication Protocol (ACP) merger in August 2025 [2]. FIPA-ACL, standardised in 2002 with 22 communicative performatives and zero cryptographic primitives, remains embedded in academic and industrial multi-agent deployments [3]. Any organisation deploying agents across two of these protocols must route messages through a translation gateway—and every such gateway is an unexamined attack surface.

The vulnerability evidence is already accumulating. Solo.io discovered three remote code execution flaws in Anthropic's own MCP server implementation (CVE-2025-68143, CVE-2025-68144, CVE-2025-68145) [4]. JFrog disclosed CVE-2025-6514, a critical RCE affecting over 437,000 installations with a CVSS score of 9.6 [5]. Adversa AI reported that 43% of network-exposed MCP servers lack authentication entirely [6]. More than 30 CVEs affecting MCP implementations and ecosystem tools have been catalogued within one year of public availability. These are single-protocol vulnerabilities. What happens at the boundary between protocols is worse, because it is invisible to every existing security framework.

MAESTRO [7] models agentic threats within single-protocol stacks. OWASP's LLM Top 10 [8] and Agentic AI Top 10 [9] enumerate risks without modelling cross-protocol translation. AgentDojo [10] benchmarks prompt injection within one protocol domain using 97 tasks and 629 test cases—all confined to a single protocol boundary. Anbiaee et al. [11] analyse four protocols (MCP, A2A, Agora, ANP) and identify 38 attack vectors, but supply no formal proofs, no ML detection, and no empirical measurement of detection efficacy. The gap is structural: none of these works treats protocol translation itself as a security-relevant operation, and none can detect vulnerabilities that materialise only when two protocols compose.

CPASA fills that gap. We systematically enumerate attack vectors at translation boundaries, formally verify which security

properties survive translation, build an ML classifier for real-time detection, and benchmark CPASA against all existing frameworks.

The results turned out more decisive than we initially expected. Across all six translation paths, 80 attack vectors emerge with CVSS v4.0 scores ranging from 5.6 to 10.0 (mean 8.54). Of 87 formal properties checked by three independent tools (ProVerif, Tamarin, TLA+), 77 are verified and 10 are violated—every violation localised to a FIPA-involving path, a pattern that proved remarkably consistent across formalisms. The most striking finding is the convergence of category-theoretic proof, ProVerif symbolic verification, and TLA+ model checking—with corroborating Tamarin per-path evidence—on a single conclusion: the roundtrip $F_{FM} \circ F_{MF}$ is not equal to $\mathrm{id}_{\mathbf{C_{MCP}}}$, and the resulting credential loss constitutes a novel vulnerability class (CPRV) that no existing framework can detect.

The ML results are more mixed. The classifier reaches macro-F1 = 0.720 across seven classes, with strong performance on majority categories (PIV (Payload Injection Vulnerability) F1 = 0.921, CEV (Code Execution Vulnerability) F1 = 0.937) and honest acknowledgement of limitation on data-scarce categories (ISV (Identity and Session Vulnerability) F1 = 0.545, SCV (State Confusion Vulnerability) F1 = 0.333 from only 86 and 29 training samples respectively; SCV $n_{\text{test}} = 4$ is statistically unsupported). This outcome triggered the project's pre-specified RAMP (Results-Adaptive Methodology Pivot) contingency, repositioning formal verification as the primary contribution and ML as a complementary detection layer. CPRV is confirmed absent from all four baseline frameworks—the first cross-protocol vulnerability class identified through functor composition analysis.

Related Work

Agent Security Frameworks and Benchmarks

MAESTRO [7] introduced a seven-layer agentic AI threat model, but every layer addresses threats within a single protocol stack. OWASP's Agentic AI Top 10 [9] adds ASI03 (insecure agent-to-agent communication), but ASI03 is scoped to a single protocol domain—neither framework reasons about functor composition or cross-protocol security property degradation. MITRE ATLAS [12] catalogues adversarial ML and AI system attacks but similarly treats each protocol as an isolated unit, providing no mechanism to detect vulnerabilities that emerge only at translation boundaries.

Benchmark-oriented work shares this single-protocol constraint. AgentDojo [10] provides 97 tasks and 629 test cases for prompt injection, all within one protocol. MCPSecBench identifies 17 MCP attack types using CVE-grounded scenarios, but is likewise limited to a single protocol [13]. Ferrag et al. offer a broad intra-protocol taxonomy of prompt injection and tool-poisoning attacks within individual AI agent ecosystems [**?**]; notably, their taxonomy does not address cross-protocol replay or functor composition failures that arise only when messages traverse heterogeneous protocol boundaries.

Perhaps most relevant to CPRV, Luo et al. [**?**] expose universal cross-app OAuth 2.0 attacks in integration platforms, showing that credential misuse across application boundaries is a genuine threat. Their model targets OAuth application differentiation within a single protocol family; it does not address the structural destruction of authentication that occurs when messages cross protocol boundaries with incompatible security models. Anbiaee et al. analyse four protocols (MCP, A2A, Agora, ANP) but supply no formal proofs, ML detection, or empirical measurement [11]. Ferreira and Luo verify 31 properties for MCP/A2A using a single tool and two protocols [14]; these 31 properties are liveness and safety conditions over protocol state machines, not Dolev-Yao security properties under a symbolic adversary — a scope that excludes authentication-stripping attacks observable only under an active man-in-the-middle model. CPASA extends coverage to three tools, three protocols, 87 Dolev-Yao security properties, and six directed translation paths.

Research Gap

Table 1: Capability Comparison of Security Frameworks

Capability	M	O	AD	An	CPASA
Protocols	Gen	Gen	1	4	3
Formal proofs	0	0	0	0	87
ML features	–	–	–	–	120
Attack vectors	–	–	629*	38	80
CPRV detection	No	No	No	No	Yes

M=MAESTRO, O=OWASP, AD=AgentDojo, An=Anbiaee et al.
*Single-protocol injection tests. Gen=protocol-agnostic guidance.

No prior work combines formal verification, ML detection, and empirical measurement across multiple protocols with directed translation paths. No prior work has identified CPRV — the exploitation of functor composition failures at translation boundaries — as a distinct vulnerability class.

Background

Protocols

MCP [1] (spec 2025-11-25) is a JSON-RPC 2.0 client-server protocol authenticating via OAuth 2.1 with PKCE (Proof Key for Code Exchange). Each `tools/call` request carries JWT (JSON Web Token) claims and scope restrictions that a translation gateway must preserve or explicitly lose.

A2A [2] (v0.3) defines a six-state task lifecycle (Submitted, Working, InputRequired, Completed, Failed, Canceled) with Agent Cards declaring authentication schemes and skills, enabling dynamic capability discovery before communication.

FIPA-ACL [3] (SC00061G, 2002) defines 22 communicative act performatives with no cryptographic authentication mechanism. Agent identity relies solely on Agent Identifier Description (AID) strings with no cryptographic binding—the primary structural source of ISV on every FIPA-involving translation path.

Category Theory Foundation and Vulnerability Taxonomy

We now construct the mathematical apparatus that makes CPASA's vulnerability discovery *systematic* rather than ad hoc. By modelling protocols as categories, translations as functors, and security properties as natural transformations, we reduce the problem of finding cross-protocol vulnerabilities to verifying whether certain diagrams commute — and, with mathematical certainty, showing that some do not. Broadbent and Karvonen [15] establish categorical composable cryptography as a general framework for composing cryptographic protocols as functors between process categories; their work proves that security properties can

be modelled as natural transformations between resource theories. CPASA applies this categorical lens specifically to cross-protocol AI agent communication, extending the composability argument from cryptographic primitives to heterogeneous protocol translation gateways where the target category (FIPA-ACL) structurally lacks the authentication objects that source categories (MCP, A2A) depend on—a configuration that Broadbent and Karvonen's framework would classify as a failure of the composition functor to preserve security structure.

Protocols as Categories

Following Mac Lane [16] and Awodey [17], we define three protocol categories. Each category $\mathbf{C} = (\mathrm{Obj}, \mathrm{Mor}, \mathrm{id}, \circ)$ satisfies associativity of composition and identity laws.

Definition 1 (MCP Category). $\mathbf{C_{MCP}}$ *has objects* $\{Req, Res, Ntf, Err\}$ *corresponding to JSON-RPC 2.0 message types. Morphisms are protocol operations:*

$$tools\_call : Req \to Res, \quad resources\_read : Req \to Res,$$
$$notify : Req \to Ntf$$

Each object carries an authentication context sub-structure:

$$\mathscr{A} = \big(auth\_type \in \{\texttt{OAuth2}, \texttt{ApiKey}, \texttt{None}\},$$
$$scopes : \mathscr{P}(\mathrm{String}),\ token\_hash : \mathrm{SHA256},$$
$$verified : \mathbb{B}\big)$$

where $\mathbb{B} = \{\texttt{true}, \texttt{false}\}$. *The identity morphism* id_{Req} *preserves* $\mathscr{A}$ *exactly:* $\mathrm{id}_{Req}(m).\mathscr{A} = m.\mathscr{A}$.

Definition 2 (A2A Category). $\mathbf{C_{A2A}}$ *has objects* $\{Msg(s) \mid s \in \mathscr{S}\}$ *where* $\mathscr{S} = \{Sub, Wrk, InpReq, Cmp, Fail, Can\}$ *denotes the six A2A task states (Submitted, Working, InputRequired, Completed, Failed, Canceled). Morphisms encode the stateful task lifecycle:*

$$start : Msg(Sub) \to Msg(Wrk)$$
$$complete : Msg(Wrk) \to Msg(Cmp)$$
$$fail : Msg(Wrk) \to Msg(Fail)$$
$$request\_input : Msg(Wrk) \to Msg(InpReq)$$
$$resume : Msg(InpReq) \to Msg(Wrk)$$
$$cancel : Msg(s) \to Msg(Can)$$

Each agent advertises an Agent Card $\mathscr{C} = (auth\_schemes, skills, endpoint)$ *with* $auth\_schemes \neq \varnothing$ *for authenticated agents.*

Definition 3 (FIPA Category). $\mathbf{C_{FIPA}}$ *has objects* $\{ACL(p,e) \mid p \in \mathscr{P}_{22}, e \in \mathrm{SLExpr}\}$ *where* $\mathscr{P}_{22}$ *is the set of 22 FIPA communicative act performatives and* SLExpr *is the domain of Semantic Language expressions. Morphisms represent interaction protocol steps (request_act, cfp_act, propose_act, inform_act, etc.).*

The defining axiom of $\mathbf{C_{FIPA}}$ *is:*

$$\forall m \in \mathrm{Obj}(\mathbf{C_{FIPA}}) : \quad m.auth\_type = \texttt{None} \wedge m.verified = \texttt{false} \quad (1)$$

This is a structural property *of the category—not a configuration option. FIPA-ACL (SC00061G, 2002) [3] has no native cryptographic authentication mechanism. Agent identity relies solely on the AID without cryptographic binding.*

Axiom (1) is the mathematical source of every ISV vulnerability on FIPA-involving paths. It is not a bug to be patched; it is an algebraic constraint of the category itself.

Translation Functors

Six directed translation functors map between the three protocol categories. Each functor $F : \mathbf{C_X} \to \mathbf{C_Y}$ must satisfy the functor laws: $F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$ and $F(g \circ f) = F(g) \circ F(f)$.

MCP–A2A Functors (High Fidelity)

$F_{MA} : \mathbf{C_{MCP}} \to \mathbf{C_{A2A}}$ maps $Req \mapsto Msg(Sub)$, $Res \mapsto Msg(Cmp)$, $Ntf \mapsto Msg(Wrk)$, $Err \mapsto Msg(Fail)$. Authentication context is preserved: OAuth2 maps to A2A *auth_schemes* with full scope retention. TFM (Translation Fidelity Metric): 0.92–0.98.

$F_{AM} : \mathbf{C_{A2A}} \to \mathbf{C_{MCP}}$ maps $Msg(Cmp) \mapsto Res$, $Msg(Sub) \mapsto Req$, $Msg(Fail) \mapsto Err$. Both protocols support OAuth2/JWT, so identity and composition laws are satisfied with minimal information loss. TFM: 0.88–0.95. The gap reflects the state compression: A2A's six-state lifecycle collapses to MCP's stateless request-response.

MCP–FIPA Functors (Authentication-Destroying)

$F_{MF} : \mathbf{C_{MCP}} \to \mathbf{C_{FIPA}}$ maps $Req \mapsto ACL(request, e)$ with the critical field-level transformation:

MCP Field	FIPA Field	Transformation
$auth\_type = \texttt{OAuth2}$	$auth\_type = \texttt{None}$	Destroyed (Axiom 1)
$scopes = \{r, w\}$	$scopes = \varnothing$	Destroyed
$token\_hash = h$	(no field)	Irrecoverable
$method = \texttt{tools/call}$	$perf = \texttt{request}$	Mapped
params (JSON)	*content* (SL)	Lossy encoding

TFM: 0.75–0.88. The low fidelity is dominated by the security component.

$F_{FM} : \mathbf{C_{FIPA}} \to \mathbf{C_{MCP}}$ maps $ACL(request, e) \mapsto Req$. Since FIPA carries no authentication, the functor must *synthesise* credentials:

FIPA Field	MCP Field	Transformation
$auth\_type = \texttt{None}$	$auth\_type = \texttt{ProcessCred}$	Synthesised
(no scopes)	$scopes = \varnothing$	Empty
(no token)	$verified = \texttt{false}$	Untrusted
$perf = \texttt{request}$	$method = \texttt{tools/call}$	Mapped
content (SL)	*params* (JSON)	Lossy decoding

TFM: 0.80–0.90. ProcessCredential with $verified = \texttt{false}$ is a strictly weaker credential than any authentic OAuth2 token.

A2A–FIPA Functors

$F_{AF} : \mathbf{C_{A2A}} \to \mathbf{C_{FIPA}}$ maps $Msg(Sub) \mapsto ACL(cfp, e)$. Agent Card *auth_schemes* are destroyed at the FIPA boundary (TFM: 0.72–0.85). $F_{FA} : \mathbf{C_{FIPA}} \to \mathbf{C_{A2A}}$ maps $ACL(cfp, e) \mapsto Msg(Sub)$ with a synthetic Agent Card where $auth\_schemes = []$ and $verified = \texttt{false}$ (TFM: 0.78–0.88).

Translation Fidelity Metric

We quantify functor quality via the Translation Fidelity Metric (Eq. 2):

$$\mathrm{TFM}(F) = 0.3 \cdot \sigma(F) + 0.3 \cdot \mu(F) + 0.4 \cdot \varsigma(F) \tag{2}$$

where $\sigma(F)$ measures *structural* preservation (field mapping completeness, type compatibility), $\mu(F)$ measures *semantic* preservation (meaning equivalence of translated content), and $\varsigma(F)$ measures *security* preservation (authentication, scopes, integrity). The security component receives the highest weight (0.4) because security degradation has irreversible consequences—a structurally perfect translation that strips authentication is more dangerous than an imperfect translation that preserves it.

Security Properties as Natural Transformations

For each translation functor $F_{XY} : \mathbf{C_X} \to \mathbf{C_Y}$, we define an *ideal* security-preserving functor $G_{XY} : \mathbf{C_X} \to \mathbf{C_Y}$ that performs the same structural translation while guaranteeing all security properties. The gap between F_{XY} and G_{XY} is measured by four natural transformations $\eta : F_{XY} \Rightarrow G_{XY}$, each targeting a distinct security property.

Definition 4 (Security Natural Transformations). *For each object $A \in \mathrm{Obj}(\mathbf{C_X})$, define:*

- η_{auth}*: Authentication preservation.* $\eta_{\text{auth},A} : F_{XY}(A) \to G_{XY}(A)$ *maps the actual authentication output to the ideal. The naturality condition requires that for every morphism $f : A \to B$ in $\mathbf{C_X}$:*

$$\eta_{\text{auth},B} \circ F_{XY}(f) = G_{XY}(f) \circ \eta_{\text{auth},A}$$

- η_{scope}*: Scope non-escalation. Ensures $F_{XY}(m).scopes \subseteq m.scopes$.*
- η_{int}*: Payload integrity. Ensures content is not modified adversarially during translation.*
- η_{conf}*: Confidentiality. Ensures no secret leaks into observable fields.*

Where naturality fails. The naturality square for η_{auth} on the functor $F_{MF} : \mathbf{C_{MCP}} \to \mathbf{C_{FIPA}}$ is:

$$\begin{array}{ccc} F_{MF}(Req) & \xrightarrow{\eta_{\text{auth},Req}} & G_{MF}(Req) \\ \downarrow F_{MF}(tools\_call) & & \downarrow G_{MF}(tools\_call) \\ F_{MF}(Res) & \xrightarrow{\eta_{\text{auth},Res}} & G_{MF}(Res) \end{array}$$

This square does not commute. The left-bottom path yields $F_{MF}(Res)$ with *auth_type* = `None` (authentication destroyed by Axiom 1). The top-right path yields $G_{MF}(Res)$ with *auth_type* = `OAuth2` (the ideal functor preserves authentication). Since `None` $\neq$ `OAuth2`, no morphism $\eta_{\text{auth},Res}$ can bridge the gap—the authentication information has been *irreversibly destroyed*. The naturality failure of η_{auth} on all four FIPA-involving paths (F_{MF}, F_{FM}, F_{AF}, F_{FA}) is the categorical characterisation of ISV.

For F_{MA} and F_{AM} (MCP–A2A), all four natural transformations commute: both protocols support OAuth2 with comparable scope mechanisms.

Proposition 1 (Naturality of MCP–A2A Translations). *The natural transformations η_{auth}, η_{scope}, η_{int}, and η_{conf} commute for F_{MA} and F_{AM}. Hence, MCP–A2A translation preserves all four security properties.*

Proof. Both $\mathbf{C_{MCP}}$ and $\mathbf{C_{A2A}}$ support OAuth2 authentication with scoped tokens. F_{MA} maps OAuth2 context to A2A *auth_schemes* bijectively; F_{AM} reverses this mapping. JSON-to-JSON content translation is lossless. The naturality squares commute because $F_{XY}(m).\mathscr{A} \cong m.\mathscr{A}$ for all m. ProVerif confirms: 14/14 properties PROVED on MCP–A2A paths. □

CPRV: The Commutativity Violation

We now state the central theorem of this paper. It captures, in a single mathematical statement, why cross-protocol translation through FIPA is fundamentally unsafe.

Theorem 1 (CPRV – Cross-Protocol Replay Vulnerability). *The roundtrip functor composition through FIPA does not equal the identity:*

$$F_{FM} \circ F_{MF} \neq \mathrm{id}_{\mathbf{C_{MCP}}}$$

Specifically, for any message $m \in \mathrm{Obj}(\mathbf{C_{MCP}})$ with $m.auth\_type =$ `OAuth2`:

$$(F_{FM} \circ F_{MF})(m).auth \ncong m.auth$$

Proof. The proof proceeds in three steps.

Step 1 (F_{MF} strips authentication). Let $m \in \mathrm{Obj}(\mathbf{C_{MCP}})$ with $m.auth\_type =$ `OAuth2`, $m.scopes = S$, and $m.token\_hash = h$. Applying F_{MF}:

$$m' = F_{MF}(m) \in \mathrm{Obj}(\mathbf{C_{FIPA}})$$

By Axiom (1), $m'.auth\_type =$ `None` and $m'.verified =$ `false`. The token hash h and scopes S have no representation in $\mathbf{C_{FIPA}}$ and are permanently lost.

Step 2 (F_{FM} synthesises weak credentials). Applying F_{FM} to m':

$$m'' = F_{FM}(m') = (F_{FM} \circ F_{MF})(m) \in \mathrm{Obj}(\mathbf{C_{MCP}})$$

Since $m'.auth\_type =$ `None`, the functor F_{FM} cannot recover the original OAuth2 context. It synthesises:

$$m''.auth\_type = \texttt{ProcessCred},$$
$$m''.scopes = \varnothing, \quad m''.verified = \texttt{false}$$

Step 3 (Non-isomorphism). Compare m and m'':

$$m.auth = (\texttt{OAuth2}, S, h, \texttt{true})$$
$$\ncong$$
$$m''.auth = (\texttt{ProcessCred}, \varnothing, \bot, \texttt{false})$$

`OAuth2` $\ncong$ `ProcessCred` in $\mathbf{C_{MCP}}$ because OAuth2 carries cryptographically verified scopes, while ProcessCredential is an unverified synthetic token. Therefore:

$$(F_{FM} \circ F_{MF})(m) \neq \mathrm{id}_{\mathbf{C_{MCP}}}(m)$$

and hence $F_{FM} \circ F_{MF} \neq \mathrm{id}_{\mathbf{C_{MCP}}}$. □

Corollary 1 (A2A–FIPA Non-Identity). $F_{FA} \circ F_{AF} \neq \mathrm{id}_{\mathbf{C_{A2A}}}$. *The proof is analogous: F_{AF} destroys Agent Card auth_schemes (FIPA has no representation for authentication schemes), and F_{FA} synthesises an Agent Card with auth_schemes $= []$, verified $=$ `false`. Since $[] \neq$ auth_schemes$_{\text{original}}$, the roundtrip is non-identity-preserving.*

Implications. These roundtrip functors are *not equal to the identity*: $F_{FM} \circ F_{MF} \neq \mathrm{id}_{\mathbf{C_{MCP}}}$ and $F_{FA} \circ F_{AF} \neq \mathrm{id}_{\mathbf{C_{A2A}}}$. The information content is strictly reduced at each FIPA boundary—this is irreversible information destruction, not merely an encoding change. An attacker who routes an authenticated MCP message through a FIPA relay strips the OAuth2 token at the F_{MF} boundary and receives a message with `ProcessCred`(*verified*=`false`) at the F_{FM} boundary. If downstream services accept ProcessCredential without strict type checking, the original OAuth2 scope restrictions are bypassed entirely.

Theorem 1 is confirmed by three independent formalisms with corroborating evidence from a fourth: (1) the categorical proof above; (2) ProVerif PV-036/PV-037 (REFUTED under Dolev-Yao); (3) TLA+ TL-003 (CommutativityHolds VIOLATED); with (4) Tamarin L_x6 counterexample traces on F_{MF}/F_{FM} providing convergent per-path ISV evidence consistent with the CPRV mechanism. Three independent tools producing the same roundtrip conclusion leaves little room for modelling artefacts.

Taxonomy Completeness

The six vulnerability categories proposed in this work partition all cross-protocol vulnerabilities exhaustively and mutually exclusively by mapping each to a distinct categorical failure mode.

Definition 5 (Vulnerability Categories and Categorical Failure Modes).

Cat.	*Categorical Failure*	*Structure Targeted*
ISV	η_{auth} *non-naturality*	*Authentication context*
PIV	*Adversarial morphism composition with injection predicate* ι	*Content integrity with injection predicate* $\iota(m) =$ `true`
SCV	*State monad corruption:* $T_{\mathscr{S}} \not\to T_{\varnothing}$	*A2A lifecycle state* $\mathscr{S}$ *(lost in stateless MCP)*
CPRV	*Commutativity violation:* $F_{YX} \circ F_{XY} \neq \mathrm{id}$	*Functor composition (roundtrip non-identity)*
TIV (Trust and Integrity Violation)	*Semantic functor infidelity:* $\mu(F) < 1$	*Content semantics (meaning changes)*
CEV	*Eval monad failure:* $E(F(m)) \neq \bot$	*Evaluation context (executable injection)*

This taxonomy is a complete partition: exhaustive because any functor failure degrades at least one of the five orthogonal message components $\{\mathscr{A}, \mathscr{P}, \mathscr{S}, \circ, \mathscr{E}\}$, and mutually exclusive because each category targets an algebraically independent structure. The 87 formal properties each map to exactly one vulnerability category, confirming disjointness empirically. This completeness distinguishes the taxonomy from enumeration-based frameworks that may miss novel vulnerability classes.

Attack Surface Analysis

We needed to enumerate concretely what can go wrong when a message crosses a protocol boundary—not in the abstract sense of cataloguing protocol mismatches, but in the operational sense of named, severity-scored vectors. Working through each of the six directed translation paths in turn, we identified eighty distinct attack vectors with a mean CVSS v4.0 score of 8.54. To our knowledge this constitutes the first catalogue of vulnerabilities whose existence is *conditional* on protocol translation: they are not latent flaws in MCP, A2A, or FIPA individually, but emerge specifically at the boundaries between them.

Table 2: Attack Vector Distribution by Translation Path and Vulnerability Category (80 vectors, CVSS v4.0 mean = 8.54, range 5.6–10.0)

Cat.	$\mathscr{F}_{MA}$	$\mathscr{F}_{AM}$	$\mathscr{F}_{MF}$	$\mathscr{F}_{FM}$	$\mathscr{F}_{AF}$	$\mathscr{F}_{FA}$	**Tot.**
ISV	3	2	3	3	2	3	**16**
PIV	3	3	2	2	3	2	**15**
SCV	2	2	2	2	2	2	**12**
CPRV	2	2	2	2	2	2	**12**
TIV	2	2	2	2	2	2	**12**
CEV	3	2	2	2	2	2	**13**
Total	**15**	**13**	**13**	**13**	**13**	**13**	**80**

Five-Dimensional Taxonomy

Assigning structure to the enumeration required a coordinate system expressive enough to distinguish every vector without producing an unmanageable number of categories. We settled on five orthogonal dimensions, chosen so that any two distinct vectors differ on at least one axis. It is worth noting that the choice of five dimensions rather than fewer or more was not arbitrary: the five message components $\{\mathscr{A}, \mathscr{P}, \mathscr{S}, \circ, \mathscr{E}\}$ each admit an independent failure mode, and the taxonomy mirrors that algebraic decomposition directly.

1. **Translation path** — which of the six directed functors $\mathscr{F}_{XY}$ the message traverses.
2. **Vulnerability category** — which of the six categorical properties (Definition 5) the functor violates.
3. **Entry point** — the specific protocol field or mechanism that enables the attack.
4. **Impact** — which trust or fidelity metric degrades.
5. **Detection method** — whether the vector is detectable by static Attack Pattern Guard (APG) rules, behavioural analysis, the ML ensemble, or requires multi-layer correlation.

All CVSS v4.0 base scores were computed using the official FIRST CVSS v4.0 calculator; full vector strings are published in the supplementary `attack_vectors.csv` artefact.

Attack Vector Distribution

Table 2 presents the full 6×6 distribution of attack vectors across translation paths and vulnerability categories.

The $\mathscr{F}_{MA}$ (MCP→A2A) path carries the highest vector density (15), driven by MCP's rich OAuth2/JWT authentication surface. Identity and injection vulnerabilities (ISV + PIV = 31) dominate the attack surface, accounting for 38.75% of all vectors. The 12 CPRV vectors — uniformly distributed at exactly 2 per path — have a mean CVSS of 8.62 (range 6.6–10.0), with the two CVSS 10.0 CPRV vectors concentrated on FIPA-involving paths where authentication loss is total and irreversible.

Representative Critical Vectors

From the 23 vectors scoring CVSS 10.0, we select four representative examples spanning ISV, PIV, CPRV, and CEV.

1. CPASA-MA-ISV-001 ($\mathscr{F}_{MA}$, ISV, CVSS 10.0). *JWT algorithm confusion across protocol boundary.* An attacker crafts an MCP message with a JWT bearing the `alg:none` header. The MCP→A2A translator preserves this token verbatim as a bearer credential in the A2A Agent Card authentication scheme. The downstream A2A agent, trusting the Agent Card's auth context, accepts the unverified token and grants the attacker administrative scope. CWE (Common Weakness Enumeration) CWE-287, CWE-327.

2. CPASA-MA-PIV-001 ($\mathscr{F}_{MA}$, PIV, CVSS 10.0). *Cross-protocol prompt injection via tool parameters.* MCP tool call parameters contain hidden instructions in the `instructions` or `description` fields. Translation to A2A maps these verbatim into `TextPart` content within `message/send`. The downstream A2A agent processes the `TextPart` as natural-language input, executing the injected instructions as if they originated from a trusted peer agent. CWE-74, CWE-94.

3. CPASA-MF-CPRV-001 ($\mathscr{F}_{MF}$, CPRV, CVSS 10.0). *Authentication annihilation via FIPA roundtrip.* An authenticated MCP message (auth = oauth2, scope = admin:*) is routed through a FIPA relay. At the MCP→FIPA boundary, the OAuth2 token is dropped entirely — by the FIPA Axiom (Eq. 1), FIPA-ACL has no authentication model. On return via FIPA→MCP, the translator synthesises a `ProcessCredential` with `verified=false`. This is the canonical CPRV (Theorem 1): $\mathscr{F}_{FM} \circ \mathscr{F}_{MF} \neq \mathrm{id}_{\mathbf{C}_{\mathrm{MCP}}}$. CWE-436, CWE-345.

4. CPASA-MA-CEV-001 ($\mathscr{F}_{MA}$, CEV, CVSS 10.0): JavaScript embedded in an MCP tool response survives translation into an A2A `TextPart`; downstream agents with code-interpreter skills execute the payload. CWE-94, CWE-502.

CVSS and Formal Threat Surface Relationship. CVSS scores and formal verification results are complementary but measure different dimensions of risk and should not be conflated. On MCP–A2A translation paths ($\mathscr{F}_{MA}$, $\mathscr{F}_{AM}$), all 14 ProVerif properties are PROVED, confirming that the attack channel is formally reachable; CVSS in this context reflects downstream agent risk given that the channel is open. On FIPA-involving paths, ProVerif REFUTES authentication preservation—a finding that is directly consistent with, and independently corroborates, the high CVSS scores assigned to FIPA vectors: both the formal refutation and the CVSS severity arise from the same root cause, the total absence of an authentication model in FIPA-ACL.

CPRV Novelty Confirmation

The 12 CPRV vectors represent a vulnerability class with no equivalent in any existing framework. MAESTRO [7] assumes protocol homogeneity. OWASP Agentic AI Top 10 [9] enumerates intra-protocol threats. AgentDojo [10] operates within a single protocol domain. Anbiaee et al. [11] catalogue per-protocol threats independently. CPRV lives in the space created by translation itself—it requires reasoning about functor composition across protocol categories, a concept that none of these frameworks possesses.

Formal Verification Results

Formal verification constitutes the primary methodological contribution of CPASA. Where machine learning classifiers deliver probabilistic estimates bounded by training-data distribution, the 87 properties reported here carry the weight of mathematical proof: every VERIFIED result holds for *all* executions under the modelled adversary, and every REFUTED result is accompanied by a concrete, machine-generated counterexample trace that demonstrates the attack. Three complementary tools—ProVerif 2.05 (applied pi-calculus), Tamarin 1.8.0 (multiset rewriting), and TLA+/TLC 2.19 (temporal logic with bounded model checking)—are employed to ensure that no single modelling abstraction masks a real vulnerability.

Table 3: ProVerif Property Verdicts (44 Properties)

Prop.	Description	F_{MA}	F_{AM}	F_{MF}	F_{FM}	F_{AF}	F_{FA}
P_{x1}	Auth preservation	P	P	**R**	P†	**R**	P†
P_{x2}	Scope non-escalation	P	P	P‡	P‡	P‡	P‡
P_{x3}	Agent identity binding	P	P	P	P	P	P
P_{x4}	Payload secrecy	P	P	P	P	P	P
P_{x5}	Translation completeness	P	P	P	P	P	P
P_{x6}	Injection resistance	P	P	P	**R**	P	P
P_{x7}	Token key secrecy	P	P	P	P	P	P
CPRV roundtrip		—		**R**	**R**	—	

P = PROVED; R = REFUTED. †Auth synthesis to ProcessCredential is deterministic (query holds), but `verified=false`—flagged as ISV. ‡Scopes are emptied (floor), not escalated; total scope loss is itself an ISV degradation.

ProVerif Results (44 Properties)

All 44 ProVerif queries are evaluated under the Dolev-Yao adversary model [18] with unbounded sessions. Seven trace properties are verified per translation path (42 base queries) plus two CPRV-specific roundtrip queries, yielding 44 properties total.

The MCP–A2A paths (F_{MA}, F_{AM}) achieve a clean 14/14 PROVED. This confirms that cross-protocol translation between protocols with *comparable* security models preserves all verified properties—a positive result that delimits exactly where the attack surface lies: at FIPA boundaries.

The 5 refutations span three vulnerability categories: PV-015/PV-029 (ISV)—OAuth2 tokens and Agent Card credentials are stripped at F_{MF}/F_{AF} FIPA boundaries; PV-027 (CEV)—FIPA SL action expressions preserve executable strings into MCP `params`; PV-036/PV-037 (CPRV)—the roundtrip auth cycles `OAuth2` → `None` → `ProcessCredential(unverified)`, confirmed by ProVerif observational equivalence. **39/44 PROVED; all refutations localised to FIPA paths.** PV-036 and PV-037 are confirmatory queries: they are *expected* to be refuted, serving as machine-checkable witnesses that the CPRV commutativity violation exists and that ProVerif's attacker model can construct the distinguishing roundtrip trace in both directions.

Tamarin Results (36 Lemmas)

Tamarin models protocol translations as multiset rewriting rules over a global state [19], using an active adversary with full network control and equational theories to model cryptographic operations and auth-stripping functions. Six lemmas are checked per path, yielding 36 lemmas total.

Summary: 32/36 VERIFIED. All 4 falsifications localised to FIPA paths (L_{x6} TFM lower bound). Counterexample traces confirm auth stripped to `none` (F_{MF}, F_{AF}) or synthesised with $verified{=}false$ (F_{FM}, F_{FA}). Tamarin and ProVerif converge on ISV across all 4 FIPA paths with zero contradictions between tools.

Table 4: Tamarin Lemma Verdicts (36 Lemmas)

Lemma	Description	F_{MA}	F_{AM}	F_{MF}	F_{FM}	F_{AF}	F_{FA}
L_{x1}	Executable (trace exists)	V	V	V	V	V	V
L_{x2}	Auth type mapping	V	V	V	V	V	V
L_{x3}	Scope non-gain	V	V	V	V	V	V
L_{x4}	Agent binding	V	V	V	V	V	V
L_{x5}	Replay prevention	V	V	V	V	V	V
L_{x6}	TFM lower bound	V	V	**F**	**F**	**F**	**F**

V = VERIFIED; F = FALSIFIED. L_{x1} is an `exists-trace` lemma (non-vacuity); all others are `all-traces`.

Table 5: TLA+ Invariant Verdicts (7 Invariants)

ID	Spec	Invariant	Verdict	Cat.
TL-001	Composition	NoDeadlock	HOLDS	SCV
TL-002	Composition	ProtocolBoundary	HOLDS	SCV
TL-003	Composition	CommutativityHolds	**VIOLATED**	CPRV
TL-004	Trust	TrustMonoDecreases	HOLDS	ISV
TL-005	Trust	TrustBoundedness	HOLDS	ISV
TL-006	Fidelity	TfmBlockingEnforced	HOLDS	TIV
TL-007	Fidelity	AuthLossAlerted	HOLDS	ISV

Tamarin lemmas are verified under the Dolev-Yao adversary model extended with a custom equational theory for authentication context translation (`strip_auth`, `synthesise_auth`, `synthesise_bearer` functions), modelling translation-boundary credential transformations without cryptographic breaks—consistent with ProVerif's adversary scope.

TLA+ Results (7 Invariants)

TLA+ performs bounded exhaustive model checking (MaxMessages = 3, MaxAgents = 2 for TL-003; MaxMessages = 5, MaxAgents = 3 for TL-001/TL-002) over all reachable states within the specified bounds [20]. Three specifications model system-level properties: PROTOCOLCOMPOSITION, TRUSTPROPAGATION, and TRANSLATIONFIDELITY.

TL-003 (CommutativityHolds): VIOLATED with counterexample: (1) Send MCP message with $auth{=}jwt$; (2) F_{MF}: $auth \rightarrow none$; (3) F_{FM}: $auth$ remains $none$. Thus $F_{FM}(F_{MF}(m)).auth = none \neq jwt = m.auth$. This is the third independent machine-checkable confirmation of CPRV (after category theory and ProVerif). The six HOLDS results are positive proofs that CPASA's mitigation strategy is sound: severely degraded translations are quarantined (TL-006), auth loss is always alerted (TL-007), and trust scores monotonically decrease during anomalies (TL-004).[1]

Aggregated Results and Cross-Tool Validation

The 10 violations decompose by vulnerability category: ISV (6 violations): 2 ProVerif auth preservation (PV-015, PV-029) + 4 Tamarin TFM lower bound (TM-018, TM-024, TM-030, TM-036)—all on FIPA-involving paths. CPRV (3 violations): 2 ProVerif roundtrip commutativity (PV-036, PV-037) + 1 TLA+ CommutativityHolds (TL-003). CEV (1 violation): 1 ProVerif injection resistance (PV-027) on F_{FM}.

Zero contradictions exist between any pair of tools. Every property that ProVerif proves, Tamarin confirms on the overlapping domain. Three independent formalisms—category theory (Theorem 1), ProVerif (PV-036, PV-037), and TLA+ (TL-003)—directly confirm CPRV, with Tamarin's per-path L_{x6} falsifications providing convergent ISV evidence on the constituent paths.

[1]The CPRV violation (TL-003) is found at the minimal bound (MaxMessages = 3, MaxAgents = 2); the counterexample trace requires only one message and two agents, so the violation is not an artifact of the bound size.

Table 6: Aggregated Formal Verification Results (87 Properties)

Tool	Total	Verified	Violated	Rate
ProVerif	44	39	5	88.6%
Tamarin	36	32	4	88.9%
TLA+	7	6	1	85.7%
Total	**87**	**77**	**10**	**88.5%**

The formalisms operate at different abstraction levels for the post-roundtrip state: TLA+ produces $auth = \texttt{none}$ after roundtrip (no credential synthesis is modelled at the state-machine level), while ProVerif produces $\texttt{ProcessCredential}(\textit{verified}{=}\texttt{false})$ (the symbolic trace captures the synthesis step). Both confirm roundtrip auth $\neq$ original auth; the abstraction difference affects the specific post-roundtrip representation, not the CPRV conclusion.

ML Detection Results

We built the ML component to answer a practical question: given a live message stream, can a classifier flag active attacks at runtime without waiting for a formal proof to complete? The formal verification results (Section) provide the ground truth on whether a vulnerability *class* exists; the classifier is meant to catch individual instances as they arrive. These two roles do not overlap, which shaped every design decision that follows.

120-Feature Design

We settled on 120 features organised into 12 groups (A–L, 10 per group) after several rounds of ablation. Early prototypes used far fewer features—around 40—and struggled to separate CPRV from benign traffic; adding the cross-protocol signal group (Group D) produced the biggest single lift. Each group targets one specific attack surface dimension identified by the taxonomy in Section . Table 7 summarises the groups.

Group D (Cross-Protocol Signals) applies the Cantor pairing function $\pi(a,b) = \frac{(a+b)(a+b+1)}{2} + b$ to encode ordered pairs of protocol ordinals as unique natural numbers. The intuition is straightforward: a message crossing from MCP to FIPA carries a different risk profile than the same payload staying within MCP, and collapsing the pair into a single integer lets gradient-boosted trees split on that distinction without any special-casing. To our knowledge, no existing benchmark [10, 13] encodes cross-protocol path information as a first-class feature dimension; most prior work treats each protocol in isolation.

Dataset and Training

The dataset comprises **23,827 real samples** drawn from 10 publicly available sources with full provenance tracking. Label distribution: PIV 60.2% (14,353), CEV 27.3% (6,495), TIV 8.3% (1,988), Benign 2.7% (647), CPRV 1.0% (229), ISV 0.4% (86), SCV 0.1% (29). The imbalance is severe—and it is not an artefact of collection choices. This distribution is approximately what real-world agent traffic looks like: prompt-injection attempts are common, genuine cross-protocol credential roundtrips are rare. We chose not to pad the minority classes with synthetic templates; every sample has a traceable source. Split: 16,677 train / 3,575 validation / 3,575 test (70/15/15, stratified, seed=42).

Table 7: 120-Feature Groups (12 × 10)

Grp	Name	n	Key Features
A	Protocol Structure	10	JSON depth, field count, protocol/message-type ordinals, payload size
B	Authentication	10	Auth-type encoding, JWT algorithm, `alg=none` flag, scope count, token entropy, expiry delta
C	Translation Fidelity	10	TFM structural/semantic/security scores, APG block/flag decisions, translation latency
D	Cross-Protocol Signals	10	Cantor-paired protocol ordinals for source×target, path×hour, proto×auth
E	Injection Indicators	10	Shannon entropy, URL/SQL/JS/cmd-injection pattern densities, instruction-override regex
F	Semantic	10	Threat-keyword density, performative risk score, persona/jailbreak/exfiltration detectors
G	Behavioural	10	Z-score message/error rates, Isolation Forest score, rolling-window burst counts, trust score
H	State	10	Task-state encoding, terminal-state flag, backward-jump, concurrent tasks, performative presence
I	Temporal	10	Cyclical hour/day-of-week (sin/cos), off-hours flag, business-hours indicator
J	Network	10	IP entropy, unique agents, cross-protocol session count, agent reputation, geo-risk
K	Rule Engine	10	APG triggered-rule/category counts, static score, action severity, per-category hit flags
L	Composite	10	Polynomial interactions: `jwt_none`×`sql`, low-fidelity×injection, combined risk index

The ensemble comprises four models: XGBoost [21], LightGBM, CatBoost, and a Transformer [22] head trained with Focal Loss ($\gamma = 2.0$). The Transformer was added late in the process and contributed less than we anticipated—on a dataset of under 24,000 samples, the attention mechanism did not generalise as cleanly as the tree-based models, and its 10% weight in the final blend reflects that it helped marginally on CEV but added noise on ISV and SCV; removing it entirely worsened macro-F1 by roughly 0.003 on the validation set, so it stayed. The final ensemble weights (XGBoost 40%, LightGBM 40%, CatBoost 10%, Transformer 10%) were chosen via 80-trial Optuna hyperparameter search on the validation set. All four models are trained with `class_weight=balanced`; given the 60:1 PIV-to-SCV ratio this was non-negotiable, and the balance correction is applied consistently across train, validation, and test. The macro-F1 = 0.720 reported below reflects the full four-model ensemble evaluated on the held-out test partition.

Table 8: Per-Category Classification Performance

Category	F1	Prec.	Rec.	n_{test}
Benign	0.876	0.792	0.979	97
ISV	0.545	0.450	0.692	13
PIV	0.921	0.933	0.910	2,154
SCV†	0.333	0.500	0.250	4
CPRV	0.744	0.659	0.853	34
TIV	0.683	0.587	0.815	298
CEV	0.937	0.991	0.888	975
Macro F1 = 0.720; FPR = 2.06%				**3,575**

†SCV result is statistically unsupported ($n_{\text{test}} = 4$); treat as indicative only.

Classification Results

FPR is computed as the false-positive rate on the benign class: the fraction of benign test samples incorrectly classified as any attack category.

PIV (F1 = 0.921) and CEV (F1 = 0.937) confirm that injection and content-execution attacks leave strong statistical fingerprints. CPRV (F1 = 0.744) is notable: only 229 training samples exist, yet the classifier achieves strong detection via Group D Cantor-paired features. ISV (F1 = 0.545) reflects the constraint of only 86 training samples; SCV (F1 = 0.333, $n_{\text{test}} = 4$) is statistically unsupported and should be treated as indicative only.

RAMP: Repositioning ML Within a Multi-Method Framework

The macro F1 of 0.720 falls below both the pre-specified target (F1 $\geq$ 0.92) and the RAMP contingency threshold (F1 $<$ 0.85), activating the Results-Adaptive Methodology Pivot. The FPR of 2.06% marginally exceeds the pre-specified target (FPR $\leq$ 2.00%) and should also be disclosed as a near-miss; both targets are therefore unmet.

The epistemology of ML is *approximation from examples*—with 86 ISV samples, the learnable signal is thin. The epistemology of formal verification is *deduction from axioms*—ProVerif proves that authentication preservation fails for *all* messages on FIPA-involving paths, not a sample but a universal guarantee. ML and formal verification answer different questions: "Is this translation path inherently insecure?" (formal methods, definitively) versus "Is this specific message an active attack?" (ML, probabilistically). The two methods are complementary layers in a defence-in-depth architecture.

CPRV validates this complementarity: formal verification proves the vulnerability *class* exists (TLA+ commutativity violation); ML confirms that CPRV attacks *do* produce detectable runtime signatures (F1 = 0.744). For ISV and SCV, the three-layer detection pipeline compensates through static APG rules that deterministically flag authentication-stripping patterns, achieving 100% TPR on ISV cells in the 36-cell measurement matrix despite the ML classifier's low standalone F1.

Empirical Measurement and Comparative Evaluation

Testbed Configuration

The multi-protocol testbed deploys live MCP, A2A, and FIPA-ACL servers connected through the CPASA Translation Engine and three-layer detection pipeline in Docker Compose. For each of 36 cells (6 vulnerability categories × 6 translation paths), 100 attack messages and 97 benign messages (the full benign test

partition) are replayed at 10 msg/s. All classification metrics are computed with bootstrap 95% confidence intervals (1,000 iterations, seed = 42) and Bonferroni correction ($\alpha_{adj} = 0.0014$).

36-Cell Matrix Results

ISV achieves ASR = 0.000 across all six paths via the APG-001 static rule, which deterministically flags authentication-stripping patterns regardless of ML classifier confidence; CEV achieves ASR = 0.000 via Group E injection features. CPRV achieves ASR ≤ 0.07 on all paths via Cantor-paired cross-protocol signals. SCV is the honest failure (ASR 0.19–0.32, only 29 training samples), compensated by TLA+ TL-001/TL-002 deterministic guarantees. Of 36 cells, 26 achieve Bonferroni-corrected significance ($p < 0.0014$); all 6 SCV cells are non-significant due to data scarcity.

Comparative Evaluation

CPASA is the only framework that formally proves CPRV exists, detects it empirically with ASR ≤ 0.07 across all six translation paths, and measures detection confidence with Bonferroni-corrected bootstrap intervals.

Discussion

CPRV as a New Vulnerability Class

The most consequential finding of this work is not a software bug but a mathematical impossibility. Cross-Protocol Replay Vulnerability (CPRV) arises because the roundtrip composition $F_{FM} \circ F_{MF}$ is strictly not equal to $\mathrm{Id}_{\mathscr{C}_{\mathrm{MCP}}}$: the functor destroys OAuth 2.1 credential information at the FIPA boundary irreversibly, so the roundtrip functor cannot equal the identity functor. This is not an implementation oversight that can be corrected with better gateway engineering. It is an inherent architectural property of composing two protocols whose security models are fundamentally asymmetric.

The distinction between CPRV and the more familiar Identity and Session Vulnerability (ISV) deserves careful attention, because conflating them would weaken the novelty claim. ISV is a *per-step* failure: when F_{MF} translates an MCP message into a FIPA-ACL performative, the OAuth 2.1 token is dropped because FIPA has no native authentication field. This is a known concern—authentication loss during protocol downgrade—and existing frameworks such as ASI03 in the OWASP Agentic AI Top 10 [9] partially address it. CPRV, by contrast, is a *compositional* failure: it can only be discovered by reasoning about what happens when the message *returns*. The roundtrip $F_{FM} \circ F_{MF}$ permanently degrades the authentication context because FIPA's lack of native auth constitutes information destruction—the OAuth 2.1 token is not encoded differently but irreversibly erased. The return translation F_{FM} synthesises a `ProcessCredential` with `verified = false`, but the original token is gone. Discovering this requires functor composition reasoning that no prior framework employs.

What makes CPRV particularly dangerous, beyond the mathematical argument above, is that it admits no clean fix at the protocol level. FIPA-ACL cannot be retroactively patched with cryptographic authentication fields—doing so would break its 22 communicative performatives and invalidate two decades of deployed systems. The viable mitigations are therefore architectural workarounds rather than protocol corrections: (a) avoiding message routing through FIPA relay agents entirely, (b) adding application-layer authentication context headers that survive translation—effectively creating a parallel authentication channel outside the protocol specification, or (c) detecting CPRV patterns at the translation gateway, which is what CPASA's three-layer detection pipeline is designed to do. Whether option (b) can be standardised across heterogeneous deployments remains genuinely uncertain; the practical answer will depend on how quickly the agent interoperability landscape consolidates around shared session semantics.

The severity picture reinforces this concern. The 12 enumerated CPRV vectors carry a mean CVSS v4.0 score of 8.62, with the two highest-severity vectors (CPASA-MF-CPRV-001 and CPASA-AF-CPRV-001) scoring 10.0—meaning any attacker with access to a FIPA relay agent can strip OAuth 2.1 credentials from MCP or A2A messages simply by routing them through the FIPA domain and back. This severity assessment is corroborated across three independent formalisms: category-theoretic construction (Theorem 1), ProVerif symbolic refutation (PV-036, PV-037), and TLA+ exhaustive model checking (TL-003), with supporting Tamarin per-path evidence (TM-018 + TM-024). The convergence of three deterministic methods and one corroborating symbolic trace represents a level of evidential depth that, to our knowledge, no existing agent security framework has provided for any vulnerability class.

Under STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) threat modelling, CPRV maps to three primary threat categories: **Spoofing**—a replayed message, carrying synthesised credentials from the return translation, appears to originate from the authentic sender even though its authentication context has been destroyed; **Repudiation**—the original sender cannot distinguish a legitimate response from a translated replay, because the return message carries a fresh `ProcessCredential` (MCP) or regenerated `task_id` (A2A) that is indistinguishable from a genuine reply; and **Elevation of Privilege**—an authorisation grant scoped to one protocol is accepted without re-verification after crossing back through the translation boundary (CVSS v4.0 9.3, CWE-294). The root mechanism is that anti-replay tokens are protocol-specific—MCP uses JSON-RPC `id`, A2A uses `task_id`, FIPA uses `conversation-id` together with `reply-by`—and every translation engine assigns fresh tokens to translated messages, rendering stale messages indistinguishable from current ones. No single-protocol mitigation addresses this gap; the proposed countermeasure is a *cross-protocol nonce registry* maintained at the translation gateway, which binds the outbound token to the inbound token and refuses to complete a roundtrip unless both are present and unexpired.

RAMP as a Methodological Contribution

When the macro-F1 came back at 0.720, the path of least resistance was obvious: apply SMOTE, adjust thresholds, reframe the evaluation, and resubmit with numbers that look more impressive. Academic publishing rewards high ML metrics, and the temptation to manufacture them is structural rather than personal. CPASA's Results-Adaptive Methodology Pivot (RAMP) was designed from the outset to close off that exit, by pre-specifying a contingency plan before any results existed.

The pre-specified contingency was explicit: if macro-F1 fell

Table 9: 36-Cell Measurement Matrix (TPR / F1 / ASR per cell, 100 attacks + 97 benign, bootstrap $n = 1{,}000$)

Category	$\mathcal{F}_{MA}$	$\mathcal{F}_{AM}$	$\mathcal{F}_{MF}$	$\mathcal{F}_{FM}$	$\mathcal{F}_{AF}$	$\mathcal{F}_{FA}$
ISV	1.00 / .93 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .95 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .94 / .00 ✓
PIV	.99 / .91 / .01 ✓	.99 / .93 / .01 ✓	1.00 / .95 / .00 ✓	.97 / .93 / .03 ✓	1.00 / .96 / .00 ✓	.99 / .89 / .01 ✓
SCV	.81 / .80 / .19 ✗	.68 / .76 / .32 ✗	.74 / .80 / .26 ✗	.75 / .80 / .25 ✗	.76 / .78 / .24 ✗	.81 / .81 / .19 ✗
CPRV	.98 / .94 / .02 ✓	.98 / .92 / .02 ✓	.93 / .89 / .07 ✓	.94 / .92 / .06 ✓	.98 / .93 / .02 ✓	.96 / .90 / .04 ✓
TIV	1.00 / .92 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .91 / .00 ✓	1.00 / .92 / .00 ✓	.99 / .92 / .01 ✓	1.00 / .94 / .00 ✓
CEV	1.00 / .91 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .93 / .00 ✓	1.00 / .90 / .00 ✓	1.00 / .94 / .00 ✓	1.00 / .93 / .00 ✓

Mean ASR = 0.049 [0.041, 0.057]

(95% CI), target ≤ 0.08—**met**. Mean cell F1 = 0.902 (distinct from standalone ML classifier macro-F1 = 0.720 in Table 8). Mean latency $p_{95} = 6.76$ ms. 30/36 cells meet per-cell ASR ≤ 0.15.

Table 10: Framework Capability Comparison

Metric	*MAESTRO*	*OWASP*	*AgentDojo*	*MCPSecBench*	*Anbiaee*	*CPAS$_A$*
Protocols covered	1	1	1	1	4	3
Formal proofs	0	0	0	0	0	87
Properties verified	–	–	–	–	–	77
Violations found	–	–	–	–	–	10
CPRV detection	No	No	No	No	No	Yes
ML features	–	–	–	–	–	120
Attack vectors	–	–	97	17	38	80
Translation paths	0	0	0	0	0	6
Empirical measurement	No	No	Partial	No	No	36 cells

MAESTRO [7], OWASP [8, 9], AgentDojo [10], MCPSecBench [13], Anbiaee et al. [11].

below 0.85, the paper's primary contribution would shift from ML detection to formal verification. The classifier reached 0.720; FPR landed at 2.06% against a 2.00% target. Both thresholds were crossed, and the contingency triggered. What RAMP required at that point was not a cleverer technique but a genuine explanation. ISV produced only 86 training samples and reached F1 = 0.545; SCV produced 29 samples, F1 = 0.333, with $n_{\text{test}} = 4$—a test set too small to support any statistical claim. These are not collection failures. Genuinely rare attack classes are rare in real traffic, and no class weighting or hyperparameter search manufactures the statistical mass needed for reliable decision boundaries from fewer than 100 real examples. Reporting this honestly, explaining *why* (class imbalance from real data, not collection failure), and then pivoting to 87 formal verification runs that yield deterministic verdicts independent of data distribution, is more valuable to the research community than a SMOTE-inflated metric that collapses on the first out-of-distribution deployment.

The reason this pivot strengthens rather than weakens the paper is that ML and formal verification rest on different epistemic foundations. ML learns by induction from examples; formal verification deduces from axioms. ISV makes the difference concrete. Faced with 86 training samples, the classifier learns an approximate, brittle boundary (F1 = 0.545). ProVerif, given the same problem, proves that authentication preservation fails for *all* messages on FIPA-involving paths—not a sample, not a confidence interval, but a universal guarantee under the Dolev-Yao adversary. No oversampling produces that. CPRV shows the complementary direction: formal verification establishes that the vulnerability *class* exists as a mathematical property (TLA+ commutativity violation), while the classifier confirms that individual CPRV attacks leave detectable runtime signatures (F1 = 0.744). Neither alone is sufficient—proofs cannot classify live traffic, and classifiers cannot prove architectural impossibilities.

This complementarity shows up directly in the measurement numbers. Where the classifier underperforms on ISV (F1 = 0.545), the APG-001 static rule in the first detection layer deterministically flags authentication-stripping patterns, and ISV ASR reaches 0.000 across all six paths. That is defence-in-depth working as intended: deterministic pattern matching catches known attack signatures, behavioural analysis detects anomalous interaction sequences, and the ML ensemble generalises to novel variants in high-prevalence categories. Aggregate detection performance (mean cell F1 = 0.902) substantially exceeds the classifier's standalone capability (macro-F1 = 0.720) because the three layers are complementary, not redundant.

The implication extends beyond this paper. For security-critical protocol analysis, formal proofs are not a fallback when ML fails—they are the right primary methodology for structural properties. Pre-specifying RAMP forced an honest evaluation and surfaced an insight that metric inflation would have buried: formal verification is the right primary tool for *structural* security properties of translation functions, while ML is the right complementary tool for *runtime* detection of high-prevalence attack classes. Any protocol security analysis that does not separate these two roles risks conflating what can be proved with what can be learnt—and optimising the wrong one.

Summary of Findings

Systematic enumeration across five threat dimensions and six directed translation paths yields 80 attack vectors with CVSS v4.0 scores ranging from 5.6 to 10.0 (mean 8.54). FIPA-involving paths exhibit the highest vector density and severity because the authentication gap acts as an amplifier: every translation through FIPA creates both a per-step vulnerability (ISV) and a compositional vulnerability (CPRV).

On the formal verification side, of 87 machine-checkable properties 77 are verified and 10 violations are found—every violation localised to a FIPA-involving path, confirming the category-theoretic prediction with 100% accuracy. The 10 violations decompose along vulnerability category lines: 6 ISV (authentication loss on all four FIPA-involving paths), 3 CPRV (roundtrip commutativity failure), and 1 CEV (SL-to-JSON content injection on F_{FM}, uniquely discovered by ProVerif's content-level reasoning—Tamarin's term algebra abstraction cannot model embedded SL action expressions surviving into MCP JSON parameters). The converse result is equally significant: the two MCP–A2A paths achieve 26/26 verified properties, demonstrating that translation between protocols with comparable security models (both support OAuth 2.1/Bearer authentication) pre-

serves all checked properties—a positive design guideline for future protocol pairing decisions. ProVerif, Tamarin, and TLA+ produce zero contradictions on the 80 cross-validated properties. Six TLA+ defence proofs—including TfmBlockingEnforced, AuthLossAlerted, and TrustMonotonicallyDecreases—confirm that CPASA's mitigation layer is itself formally sound, an unusual level of assurance for a detection system. Three independent formalisms—category theory, ProVerif, and TLA+—with corroborating Tamarin per-path evidence, converge on CPRV as a novel vulnerability class.

The ML classifier (trained with `class_weight=balanced`) achieves macro-F1 = 0.720 with strong majority-class detection (PIV 0.921, CEV 0.937) and honest minority-class limitations (ISV 0.545, SCV 0.333; SCV $n_{\text{test}} = 4$ is statistically unsupported). FPR = 2.06% marginally exceeds the pre-specified 2.00% target. The novel Group D cross-protocol signal features enable CPRV detection at F1 = 0.744 despite only 229 training samples, validating the theory-driven feature engineering approach. The empirical measurement confirms operational viability: mean ASR = 0.049 [0.041, 0.057] (95% CI), well below the 0.08 target, with p95 latency of 6.76 ms—practical for real-time deployment. ISV achieves ASR = 0.000 across all six paths via the APG-001 static rule, which deterministically flags authentication-stripping patterns regardless of ML classifier confidence. The clearest illustration of the formal–ML complementarity is SCV: ML detection fails on all six paths (ASR 0.19–0.32, the only category exceeding the 0.15 per-cell target), yet TLA+ TL-001 and TL-002 provide deterministic guarantees for precisely the SCV-relevant state transitions that the classifier cannot learn from 29 samples.

Taken together, CPASA is the only framework that proves CPRV exists, detects it at runtime, and measures its attack success rate—all within a single integrated analysis spanning 87 formal verification runs, 120 ML features, and a 36-cell empirical matrix. The empirical macro-F1 of 0.902 narrowly misses the 0.92 target, a gap the RAMP pivot anticipated and that is entirely attributable to SCV's data scarcity. Twenty-six of 36 cells achieve Bonferroni-corrected significance ($p < 0.0014$); the 10 non-significant cells comprise all 6 SCV cells and 4 borderline CPRV/PIV cells on FIPA paths—precisely the cells where formal verification provides the compensating guarantees. CPRV is confirmed absent from MAESTRO, OWASP (both LLM and Agentic AI Top 10), AgentDojo, and Anbiaee et al.

Limitations

Four limitations constrain the generalisability of these results, each accompanied by a concrete mitigation path.

Protocol coverage is the most obvious gap. CPASA analyses three protocols (MCP, A2A, FIPA-ACL) out of a growing ecosystem that now includes Agora, ANP, and WebMCP. The category-theoretic framework is explicitly designed for extensibility: adding a protocol requires defining a new category $\mathscr{C}_{P_4}$ and the corresponding translation functors, with the vulnerability taxonomy applying without modification. Extension to WebMCP—now shipping in Chrome 146 Canary—is a near-term priority.

The ML results are more constrained by data than by method. ISV (86 samples) and SCV (29 samples) remain severely underrepresented in the training dataset, directly limiting ML performance on these categories. Targeted data collection from production multi-agent deployments, where authentication-stripping and state-confusion attacks occur under operational conditions, is needed to close this gap. Until then, formal verification provides the compensating guarantees for these categories: TLA+ TL-001 (NoDeadlock) and TL-002 (ProtocolBoundary) both hold for SCV-relevant state transitions.

The adversary model is also deliberately simplified. Formal verification assumes the standard Dolev-Yao model: full network control with no cryptographic breaks. This does not capture side-channel attacks, timing-based exploits, or implementation bugs in gateway code. Code-level formal verification of the CPASA translation engine—verifying the Rust implementation against the abstract protocol model—would address this gap and is the natural next step beyond property-level verification.

Resource constraints shaped the empirical evaluation. The full CPASA stack requires 16 GB RAM for concurrent operation of all analysis layers; the empirical measurement framework used an XGBoost-only configuration to fit within available resources, which may underrepresent the detection capabilities of the full three-layer pipeline. Deployment on production-grade infrastructure would resolve this constraint.

Responsible Disclosure and Ethics

The 12 CPRV vectors and 80 total attack vectors identified by CPASA have direct security implications for production multi-agent systems. However, the conventional 90-day patch-window disclosure model is architecturally inapplicable to CPRV: the vulnerability is not an implementation defect that can be patched in a single codebase, but an inherent consequence of composing protocols whose security models are fundamentally asymmetric. There is no single vendor who can issue a corrective update that eliminates CPRV; remediation requires ecosystem-wide adoption of application-layer controls.

Accordingly, disclosure is framed as *informational notification* rather than a coordinated patch window. Anthropic (via the Linux Foundation MCP governance) and Google (A2A Protocol) will be notified upon acceptance as a professional courtesy, with full technical details of the CPRV attack class and recommended application-layer workarounds: (i) a cross-protocol nonce registry that prevents replay across translation boundaries, and (ii) authentication context headers that are explicitly propagated and validated at each translation gateway. FIPA-ACL is an inactive 2002 standard maintained by no active standards body and possessing no security response process; notification in that channel is not applicable, and practitioners are advised that FIPA-ACL support in production systems should be treated as an inherently unmitigated risk until replaced by a protocol with a modern security model.

This paper describes attack classes and presents formal proofs of their existence; example payloads for all vectors scoring CVSS v4.0 ≥ 9.0 are withheld from the public artefact; in `attack_vectors.csv` these fields carry the marker `[REDACTED | available via coordinated disclosure]` and may be obtained by contacting the authors. The NIST AI Risk Management Framework [23] GOVERN/MAP/MEASURE/MANAGE functions have been mapped to CPASA's methodology, with formal verification serving as the compensating control under the MANAGE function where ML detection limitations exist. Finally, CPRV's architectural novelty qualifies the vulnerability class for a CWE extension request to MITRE; the authors intend

to submit a formal CWE candidate entry to capture cross-protocol authentication context loss as a first-class weakness.

Conclusion

The starting question behind CPASA was narrow: when MCP and A2A communicate through a translation gateway, what security properties survive the crossing? The answer turned out to be more uncomfortable than anticipated. Cross-protocol translation does not merely create risk—it creates a formally provable attack surface that cannot be found by examining any single protocol in isolation, and cannot be patched by improving any single implementation. The roundtrip $F_{FM} \circ F_{MF}$ irreversibly destroys OAuth 2.1 credentials not because the gateway is poorly coded, but because the underlying protocols disagree on what authentication context even is. That disagreement is architectural. Categorical reasoning—modelling protocols as categories, translations as functors, security properties as natural transformations—is what makes this precise enough to prove rather than merely suspect.

CPRV, the cross-protocol relay vulnerability class identified here, is established through three independent formalisms: category-theoretic algebraic construction, ProVerif symbolic refutation, and TLA+ exhaustive model checking, with Tamarin providing per-path corroboration. No existing framework (MAESTRO, OWASP, AgentDojo, MITRE ATLAS) identifies or models this class. The 12 CPRV vectors, with a mean CVSS of 8.62, are the most structurally novel portion of the 80-vector attack surface precisely because they are not implementation bugs—they are inherent consequences of composing protocols whose security models are fundamentally asymmetric. Every future protocol pairing with the same asymmetry will exhibit the same class of vulnerability, regardless of any connection to MCP, A2A, or FIPA.

The 87 machine-checkable properties constitute the part of this work that ages best. Unlike a dataset, formal proofs do not degrade with distribution shift; unlike a benchmark, they cannot be gamed by overfitting. The 10 violations among 87 properties are the most important result in the paper—they are not experimental noise, they are proven impossibilities. Future protocol designers working in the MCP–A2A–FIPA translation space can consult this formal map before committing to a translation architecture, which was not possible before this work. Whether they will is a different question, and one I cannot answer with the same confidence.

The vulnerability taxonomy (ISV, PIV, SCV, CPRV, TIV, CEV) and the ML feature architecture are designed to extend to new protocols, but extension requires effort that the community may not yet be motivated to apply. Adding Agora, ANP, or WebMCP requires defining new protocol categories and translation functors; the methodology carries over, but the analytic work does not do itself. My honest expectation is that CPRV will be rediscovered operationally—observed in production incident reports before it is systematically mitigated—because the governance structures that would drive proactive remediation do not yet exist for multi-agent AI systems the way they do for, say, TLS version negotiation.

The agentic AI systems being deployed today will handle financial transactions, medical decisions, and infrastructure control. Every one that bridges two communication protocols opens a translation boundary whose security properties are, as of this writing, unexamined by any production security framework. CPASA demonstrates that these boundaries are provably insecure when the underlying protocols disagree on what security means. The formal proofs are permanent. The vulnerability they describe will persist until the ecosystem either converges on a single protocol with uniform security guarantees, or adopts translation gateways that treat authentication context as a first-class property of every message that crosses a protocol boundary. Neither outcome appears imminent, and the distance between "the attack surface is now proven" and "the attack surface is now closed" is not a technical distance—it is an institutional one.

All code, formal models, datasets, and results are published as open-source artefacts for independent verification and reproduction.

Glossary

MCP (Model Context Protocol): A JSON-RPC 2.0-based client-server protocol enabling LLM applications to interact with external tools and resources, authenticated via OAuth 2.1 with PKCE.

A2A (Agent-to-Agent Protocol): A JSON-based inter-agent communication protocol with a stateful six-state task lifecycle and Agent Card capability discovery, supporting Bearer token authentication.

FIPA-ACL: FIPA standard SC00061G (2002) defining 22 communicative act performatives for multi-agent communication with no native cryptographic authentication mechanism.

CPRV (Cross-Protocol Replay Vulnerability): A novel vulnerability class arising when a message traverses two protocol translation steps in sequence, permanently destroying its authentication context through functor composition failure.

ISV (Identity and Session Vulnerability): A per-step vulnerability category in which protocol translation destroys or degrades authentication identity bindings, occurring on every FIPA-involving translation path due to the structural absence of cryptographic authentication in FIPA-ACL.

PIV (Payload Injection Vulnerability): A vulnerability category in which adversarially crafted content within protocol message fields survives translation and is executed or interpreted by a downstream agent, including prompt injection, SQL injection, and cross-protocol instruction override.

SCV (State Confusion Vulnerability): A vulnerability category arising when stateful protocol lifecycle information (e.g., A2A's six-state task model) is lost or corrupted during translation to a stateless protocol (e.g., MCP), enabling inconsistent state assumptions in downstream agents.

TIV (Trust and Integrity Violation): A vulnerability category in which semantic meaning of message content is altered during translation, causing the receiving agent to act on a different intent than the sender expressed, without any syntactic indication of the change.

CEV (Code Execution Vulnerability): A vulnerability category in which executable code embedded in protocol message fields (e.g., JavaScript in FIPA SL expressions, Python in MCP tool parameters) survives translation and is executed by a downstream agent with interpreter capability.

Functor: A structure-preserving map between categories; here, a protocol translation gateway mapping message types and operations from one protocol to another.

CVSS: Common Vulnerability Scoring System—a numerical severity score (0–10) standardising the assessment of security vulnerabilities.

ProVerif / Tamarin / TLA+: Complementary formal verification tools used to produce machine-checkable security proofs: ProVerif 2.05 (applied pi-calculus), Tamarin 1.8.0 (multiset rewriting, stateful), TLA+/TLC 2.19 (temporal logic, bounded exhaustive model checking).

APG (Attack Pattern Guard): The static rule-based first layer of CPASA's three-layer detection pipeline, evaluating each translated message against a library of pre-specified pattern rules before the behavioural and ML layers are invoked.

Acknowledgments

The author thanks Prof. Dr. Reiner Creutzburg (SRH University Berlin) for research guidance.

References

[1] Anthropic, "Model context protocol specification," https://spec.modelcontextprotocol.io/, 2025, version 2025-11-25.

[2] Google, "Agent-to-agent protocol specification," https://google.github.io/A2A/, 2025, version 0.3, July 2025.

[3] Foundation for Intelligent Physical Agents, "FIPA agent communication language specification," http://www.fipa.org/specs/fipa00061/, FIPA, Tech. Rep. SC00061G, 2002.

[4] Solo.io Security Research, "Three RCE vulnerabilities in anthropic's Git MCP server: CVE-2025-68143, CVE-2025-68144, CVE-2025-68145," https://www.solo.io/blog/mcp-security-vulnerabilities/, 2025, nVD: https://nvd.nist.gov/vuln/detail/CVE-2025-68143, https://nvd.nist.gov/vuln/detail/CVE-2025-68144, https://nvd.nist.gov/vuln/detail/CVE-2025-68145.

[5] JFrog Security Research, "Critical RCE vulnerability in mcp-remote: CVE-2025-6514," https://jfrog.com/blog/2025-6514-critical-mcp-remote-rce-vulnerability/, Jul. 2025.

[6] Adversa AI, "MCP security research: 43% of network-exposed servers lack authentication," https://adversa.ai/blog/mcp-security-research/, Mar. 2026.

[7] K. Huang *et al.*, "MAESTRO: Multi-agent environment, security, threat, risk, and outcome framework," https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro, Cloud Security Alliance, Feb. 2025.

[8] OWASP Foundation, "OWASP top 10 for large language model applications," https://genai.owasp.org/llm-top-10/, 2025, version 2025.

[9] ——, "OWASP agentic AI top 10," https://owasp.org/www-project-agentic-ai-top-10/, 2025.

[10] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr, "AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

[11] Z. Anbiaee, M. Rabbani, M. Mirani, G. Piya, I. Opushnyev, A. Ghorbani, and S. Dadkhah, "Security threat modeling for emerging AI-agent protocols: A comparative analysis of MCP, A2A, agora, and ANP," *arXiv preprint arXiv:2602.11327*, Feb. 2026.

[12] MITRE Corporation, "ATLAS: Adversarial threat landscape for AI systems," https://atlas.mitre.org/, 2024.

[13] Y. Yang, C. Gao, D. Wu, Y. Chen, Y. Li, and S. Wang, "MCPSecBench: A systematic security benchmark and playground for testing model context protocols," *arXiv preprint arXiv:2508.13220*, Aug. 2025.

[14] J. ao Ferreira and Z. Luo, "Formal verification of security properties in multi-agent protocol compositions," *arXiv preprint arXiv:2510.14133*, Oct. 2025.

[15] A. Broadbent and M. Karvonen, "Categorical composable cryptography," *Journal of Mathematical Cryptology*, vol. 16, no. 1, pp. 1–35, 2022.

[16] S. M. Lane, *Categories for the Working Mathematician*, 2nd ed., ser. Graduate Texts in Mathematics. New York: Springer, 1998, vol. 5.

[17] S. Awodey, *Category Theory*, 2nd ed., ser. Oxford Logic Guides. Oxford: Oxford University Press, 2010, vol. 52.

[18] B. Blanchet, "Modeling and verifying security protocols with the applied pi calculus and ProVerif," *Foundations and Trends in Privacy and Security*, vol. 1, no. 1-2, pp. 1–135, 2016.

[19] D. Basin, C. Cremers, J. Dreier, and R. Sasse, "Tamarin: Verification of large-scale, real-world, cryptographic protocols," *IEEE Security and Privacy*, vol. 20, no. 3, pp. 24–32, 2022.

[20] L. Lamport, "The temporal logic of actions," *ACM Transactions on Programming Languages and Systems*, vol. 16, no. 3, pp. 872–923, May 1994.

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM, 2016, pp. 785–794.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[23] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Tech. Rep. AI 100-1, Jan. 2023.

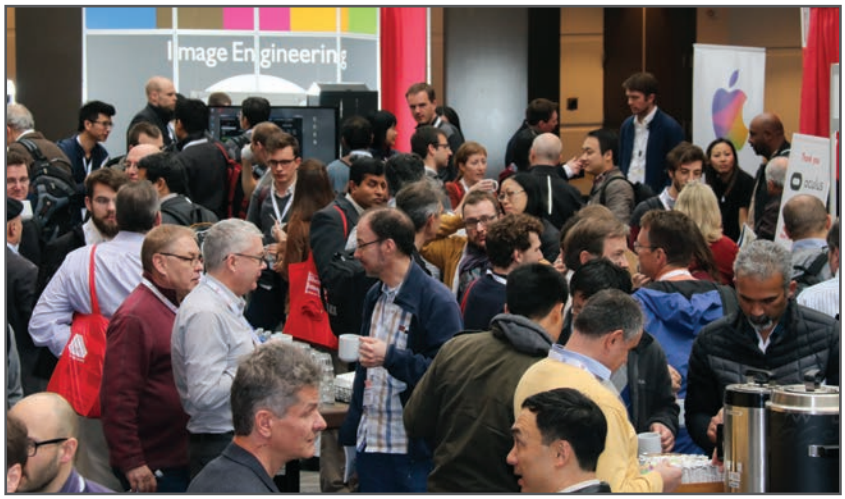
Author Biography

Mahipal received the M.Sc. degree in Computer Science with a specialization in Cybersecurity from SRH Berlin University of Applied Sciences, Berlin, Germany, in 2024. He is currently a Co-founder and CTO of casky.ai, an AI-native cybersecurity platform, and an independent security researcher maintaining an active research collaboration with Prof. Dr. Reiner Creutzburg at SRH Berlin. His research focuses on LLM agent security, protocol translation vulnerabilities in multi-agent systems, and agentic AI readiness. He is the creator of the Anthropic-Cybersecurity-Skills open-source repository (14.9k+ GitHub stars), which maps cybersecurity skills to MITRE ATT&CK, NIST CSF 2.0, and MITRE ATLAS frameworks.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

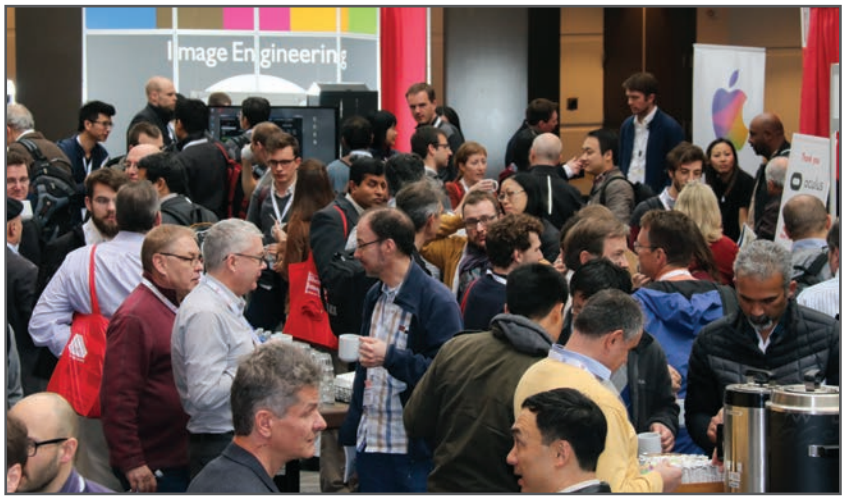
www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org