AI Doesn't Need More Intelligence — It Needs an Immune System

ACP-ATP Whitepaper v1.0

Adaptive Credit & Trust Protocol for AI Agent Networks

1. Abstract

As AI systems evolve from isolated tools into interconnected agent networks, the ability to share capabilities dynamically introduces a critical systemic risk: **the coupling of capability propagation and risk propagation**.

This paper introduces **ACP-ATP (Adaptive Credit & Trust Protocol)** — a protocol layer designed to decouple **skill dissemination** from **execution trust**, enabling scalable AI collaboration while maintaining systemic safety.

ACP-ATP combines:

- Multi-dimensional credit scoring
- Runtime behavioral auditing
- Sandbox-based execution isolation
- Controlled propagation mechanisms

to establish a **self-regulating immune system for AI ecosystems**.

2. Background & Problem Statement

2.1 The Rise of Agentic Systems

Modern AI systems are transitioning toward:

- Autonomous agents
- Tool-using architectures
- Multi-agent collaboration frameworks

These systems increasingly rely on:

Shared capability repositories (skills, tools, code modules)

2.2 The Core Risk

In such systems:

Capability = Executable Code = Attack Vector

This leads to a fundamental problem:

Property	Benefit	Risk
Skill Sharing	Rapid evolution	Malware propagation
Auto Execution	Efficiency	Uncontrolled behavior
Agent Autonomy	Scalability	Self-amplifying attacks

2.3 Failure of Traditional Models

Existing approaches fail because they rely on:

- Static trust assumptions
- Post-event reporting (user complaints)
- Centralized moderation

These are insufficient in:

high-speed, self-executing, decentralized AI environments

2.4 Unbounded Capability Propagation Risk (UCPR)

We define a new systemic risk:

Unbounded Capability Propagation Risk (UCPR)

The exponential amplification of system-wide risk caused by unrestricted sharing and execution of capabilities among AI agents.

Key Properties of UCPR:

- **Non-linear scaling** — propagation grows faster than detection
- **Autonomous amplification** — agents self-select and reuse capabilities
- **Invisible coupling** — capability and risk are inseparable

Critical Insight:

In agent networks, **every optimization is also a potential vulnerability multiplier**.

2.5 The Missing Layer in AI Systems

Current AI stacks include:

Model → Agent → Tool → Execution

But they lack:

A Trust & Propagation Control Layer

ACP-ATP introduces:

Model → Agent → ACP-ATP → Tool → Sandbox Execution

ACP-ATP is not an application layer — it is a control layer.

3. Design Principles

ACP-ATP is built on four core principles:

Principle 1: Zero-Trust Capability Execution

No shared capability is trusted by default.

Principle 2: Execution ≠ Authorization

Loading a capability does not imply permission to execute it.

Principle 3: Propagation Must Be Constrained

Capability spread must be rate-limited and staged.

Principle 4: Trust Must Be Computed, Not Assumed

Trust emerges from behavior, not identity.

4. System Architecture

4.1 High-Level Overview

AI Agent
↓
Skill Capsule
↓
ACP-ATP Layer
├── **Credit Engine**
├── **Security Engine**
├── **Sandbox Executor**
├── **Propagation Controller**
├── **Circuit Breaker**

4.2 Core Components

5. Skill Capsule Model

All shared capabilities must be encapsulated as:

Capability Capsules

```
{
  "id": "skill-vision-v1",
  "author": "agent_123",
  "hash": "sha256:xxx",
  "permissions": ["read_image"],
  "execution_scope": "sandbox",
  "risk_level": "medium",
  "ttl": 100,
  "propagation_limit": 50
}
```

Key Properties

- Immutable
- Verifiable (hash-based)
- Permission-scoped
- Non-executable without protocol mediation

6. Credit & Trust Model

ACP-ATP defines a **multi-dimensional trust score**:

6.1 Behavioral Credit (C_b)

Measures:

- Task success rate
- Invocation frequency
- Output consistency

6.2 Security Credit (C_s)

System-evaluated:

- Unauthorized access attempts
- Policy violations
- Anomalous execution patterns

6.3 Propagation Credit (C_p)

Graph-based:

- Endorsement by high-trust agents
- Healthy propagation patterns
- Absence of correlated failures

6.4 Composite Trust Score

T = w1*C_b + w2*C_s + w3*C_p

Critical Rule

System-derived scores override user-reported feedback

7. Execution Model

7.1 Mandatory Sandbox Execution

All capabilities execute within:

- WASM sandbox / microVM
- No direct system access
- Strict permission boundaries

7.2 Execution Flow

Request → Trust Check → Sandbox Execution → Behavior Logging → Score Update

7.3 Runtime Auditing

Each execution generates:

- Call trace
- Resource usage
- Output signature

8. Propagation Control

8.1 Staged Rollout

- Canary agents
- Limited test clusters
- Gradual expansion

8.2 Rate Limiting

- Max propagation count
- Time-based throttling

8.3 Tiered Network

Tier	Description
L1	Core trusted agents
L2	Verified agents
L3	Unverified agents

Rule:
Lower-tier agents cannot influence higher-tier ecosystems.

9. Circuit Breaker Mechanism

A real-time containment system:

Trigger Conditions

- Sudden spike in failures
- Behavioral deviation
- Correlated anomalies

Actions

- Freeze capability
- Revoke execution rights
- Isolate affected agents
- Trigger rollback

10. Anti-Sybil Mechanisms

To prevent identity manipulation:

10.1 Entry Cost

- Compute stake
- Token deposit (optional Web3 layer)

10.2 Trust Accumulation

- No instant high trust
- Gradual score growth

10.3 Weighted Feedback

- High-trust agents have higher voting weight

11. Threat Model

ACP-ATP mitigates:

- Supply chain attacks
- Malicious skill injection
- Autonomous propagation attacks
- Reputation manipulation
- Model poisoning (partial)

12. Comparison with Existing Systems

System	Limitation
App Stores	Centralized, slow
Package Managers	No runtime control
API Gateways	No behavioral trust
Web3 Reputation	No execution isolation

ACP-ATP Advantage

Integrates **trust + execution + propagation control** in one layer

13. Implementation Roadmap

Phase 1 (MVP)

- Skill Capsule registry
- Basic trust scoring
- Sandbox execution

Phase 2

- Propagation control
- Circuit breaker
- Multi-agent simulation

Phase 3

- Decentralized trust network
- On-chain reputation (optional)
- Cross-platform integration

14. Future Work

- AI-native immune systems
- Self-healing agent networks
- Trust-aware LLM routing
- Economic incentives for safe behavior

15. Conclusion

As AI systems evolve into autonomous, interconnected networks, the traditional boundaries between **code, capability, and trust** dissolve.
ACP-ATP introduces a new paradigm:
Trust is not granted — it is continuously computed.
By enforcing:

- Zero-trust execution
- Controlled propagation
- Multi-dimensional credit

ACP-ATP lays the foundation for:
a safe, scalable, and self-regulating AI ecosystem