Swordsman and Mage: Dual Agents Derived from the First Person

Protect or Delegate → Reflect and Connect

(S ⫫ M)|FP

privacymage

0xagentprivacy

https://agentprivacy.ai

December 11, 2025

Version 4.6

"Privacy is my blade, knowledge is my spellbook."

"Agents can only promise their own behavior." — Promise Theory

Contents

1 Notation **5**

2 Terminology Note **5**

2.1 Core Terminology . . . 5

2.2 Spellbook Learning Pathway . . . 5

2.3 Cross-Document Translation . . . 6

3 Promise-Theoretic Foundations **6**

3.1 The Autonomy Axiom . . . 6

3.2 The First Person System as Superagent . . . 6

3.3 The Gap as Irreducible Promise . . . 7

3.4 Assessment and Trust . . . 7

3.5 Invitation vs. Attack . . . 8

3.6 Coordination Promises and Spells . . . 8

3.7 VRCs as Promise Bundles . . . 8

4 Why This Matters Now: Personal AI and the Comprehension Protocol **8**

4.1 The Personal AI Future Requires . . . 8

4.2 Why Dual Agents Make This Better . . . 9

5 The Relationship Proverb Protocol (RPP) **9**

5.1 The Threat Model . . . 9

5.2 The RPP Defense . . . 9

5.3 RPP as Prompt Injection: Creating Ciphers Between People . . . 9

5.4 Example . . . 10

6 Private Proverb Inscriptions **10**
6.1 Asymmetric Commitment Structure . . . 10
6.2 Social Recovery Through Understanding . . . 10
6.3 Selective Disclosure . . . 10

7 The Inflection Point **10**
7.1 Why We Must Act Now . . . 10
7.2 The Alternative Path . . . 11

8 The 7th Capital: Behavioral Data as Personal Wealth **11**
8.1 The 7th Capital: Behavioral Sovereignty . . . 11
8.2 The Thesis . . . 11

9 The Dual-Agent Architecture **11**
9.1 Agent Definitions . . . 11
9.2 The Mathematical Constraint . . . 12
9.3 Server User-Agents: Specialized Dual-Agent Marketplace . . . 12
9.4 The Custom Marketplace: Privacy Dual Agent Primitives . . . 13

10 The Reconstruction Ceiling **14**
10.1 Single-Agent Problem . . . 14
10.2 Dual-Agent Solution with Separation . . . 14

11 The Topology of Privacy: The Triangle That Cannot Collapse **14**

12 Layer 0: Verified Personhood **14**
12.1 First Person Network . . . 15
12.2 Mathematical Requirement . . . 15

13 Initial Protocol Stack **15**
13.1 Layer 1: Agent Identity . . . 15
13.2 Layer 2: Relationship Credentials . . . 15
13.2.1 How VRCs Form Through RPP . . . 15
13.2.2 Bilateral Proverb Recovery . . . 15
13.3 Layer 3: Private Value Transfer . . . 16
13.4 Layer 4: Private Communication . . . 16
13.5 Layer 5: Collective Intelligence . . . 16

14 The Economics of Trust Networks **16**
14.1 The Compression-Trust-Value Loop . . . 16
14.2 Why This Creates Economic Value . . . 16
14.3 The Economic Flywheel . . . 16

15 The Spellbook as Semantic Infrastructure **16**
15.1 Three Core Functions . . . 17
15.1.1 1. Efficiency Through 70:1 Compression Ratio . . . 17
15.1.2 2. Verification Without Surveillance . . . 17
15.1.3 3. Sybil Resistance Through Entropy . . . 17
15.2 Story Fracture with Principle Convergence . . . 17

16 Budget System: Making Privacy Tangible **17**
16.1 Swordsman Budget (C_S) . . . 17
16.2 Mage Budget (C_M) . . . 17
16.3 The Fundamental Constraint . . . 18
16.4 Progressive Trust Tiers . . . 18

17 Chronicles: Narrative as Verification Layer **18**
17.1 Unique Derivation as Verification Signal . . . 18
17.2 The 70:1 Efficiency Gain . . . 18
17.3 Emerging Marketplace for Custom Chronicle Experiences . . . 18

18 The MyTerms Swordsman **19**
18.1 Cookie Slashing . . . 19
18.2 Cursor State as Human-in-the-Loop Audit . . . 19
18.3 State Changes as MCP Integration . . . 19
18.4 Budget Monitoring . . . 19
18.5 MyTerms Negotiation . . . 19

19 Privacy as Capital: Value Multiplication Through Trust **19**
19.1 The Tier System and Multipliers . . . 19
19.2 The Compounding Effect . . . 19

20 Web of Trust Integration **19**
20.1 Trust Graph Queries with Chronicled Audit . . . 20
20.2 Compatible Trust Protocols . . . 20
20.3 Trust Graph Queries Don't Compromise Privacy . . . 20

21 Intel Pools: Collective Intelligence Without Surveillance **20**
21.1 Access Requirements . . . 20
21.2 The Selective Disclosure Principle . . . 21

22 The 7th Capital Thesis: Behavioral Sovereignty as Wealth **21**
22.1 Extraction Versus Creation . . . 21
22.2 The Value Multiplier . . . 21

23 The Tetrahedral Future: Evolution from Two to Four **21**
23.1 Functional Requirements of Sovereignty . . . 22
23.2 The Emerging Agents . . . 22
23.3 The Tetrahedral Structure . . . 22

24 Your Proverb Revisited: The VRC Complete **22**
24.1 You've Just Completed the Foundation for a VRC . . . 22
24.2 How VRCs Form Organically . . . 22

25 Document Context **22**
25.1 This Whitepaper . . . 23
25.2 The Promise Theory Reference . . . 23
25.3 The Research Paper . . . 23
25.4 The Privacymage Spellbook . . . 23
25.5 Collaborative Development . . . 23

26 The Architectural Truth **23**
26.1 The Foundation 23
26.2 The Infrastructure 24
26.3 The Economics 24
26.4 The Principle 25

27 Document Metadata **25**
27.1 Version History 26

1 Notation

This document uses two parallel notation systems:

Mathematical	Symbolic	Meaning
S	S	Swordsman agent
M	M	Mage agent
FP	FP	First Person (human)
$(Y_S \perp\!\!\!\perp Y_M)\|X$	S $\perp$ M$\|X$	Conditional independence given private state
$H(X)$	—	Entropy of private state
C_S, C_M	—	Information budgets for Swordsman and Mage
$R_{\max}$	—	Maximum reconstruction efficiency

Mathematical notation appears in formal statements; symbolic notation in narrative sections.

2 Terminology Note

This whitepaper uses precise mathematical and architectural language.

2.1 Core Terminology

Dual Agents ($S \perp\!\!\!\perp M$)
: Two agents with conditional independence. Swordsman (protection) and Mage (delegation)

First Person
: You, the human whose sovereignty is protected (capitalized throughout to emphasize agency)

Reconstruction Ceiling ($R < 1$)
: Mathematical guarantee that adversaries cannot fully reconstruct your private state from observations

Signal
: Ongoing proverb posting (0.01 ZEC each), continuous demonstration of comprehension

Genesis Ceremony
: One-time agent pair origination, 1 ZEC ($500), different from signals

Spellbook
: Source material for learning (13 Acts, plus 30 tales in Zero Spellbook)

RPP (Relationship Proverb Protocol)
: Compression protocol proving comprehension—1 proverb formed = 1 signal posted

2.2 Spellbook Learning Pathway

How narrative learning connects to infrastructure:

- Read spellbook content (Acts or tales)
- Form a proverb showing comprehension (RPP compression)
- Post signal (1 proverb = 1 signal = 0.01 ZEC)

- Build trust tier through sustained signals (50+ = Light, 150+ = Heavy, 500+ = Dragon)
- Qualify for guardian candidacy (proven reconstruction ability)

Key insight: Guardian candidates *prove* reconstruction/compression ability through demonstrated spellbook learning. Signals are proof of comprehension, not just fees.

2.3 Cross-Document Translation

This document uses **mathematical/architectural** terminology:

- Technical: $S \perp\!\!\!\perp M|X$, reconstruction ceiling $R < 1$, information-theoretic bounds
- Architecture: Dual agents, separation primitives, conditional independence

Other documents translate these concepts:

- **Spellbook:** Narrative/mythological (Soulbis, Soulbae, the Gap, Acts/Arcs)
- **Tokenomics:** Economic/practical (SWORD, MAGE, signal fees, guardian mechanics)
- **Promise Theory Reference:** Formal semantic foundations (autonomy axiom, superagent, irreducible promise)

3 Promise-Theoretic Foundations

The dual-agent architecture is a rigorous implementation of **Promise Theory** (Bergstra & Burgess, 2019), established semantics for autonomous agent coordination.

3.1 The Autonomy Axiom

> "An agent can only make promises about its own behavior. No agent can make a promise on behalf of another agent."

This is why single agents cannot resolve the privacy-delegation paradox. A single agent attempting to promise both protection AND delegation violates the autonomy axiom—it promises in domains it cannot independently control.

The dual-agent architecture enforces this axiom:

- **Swordsman** promises protection behaviors (boundaries, disclosure control)
- **Mage** promises delegation behaviors (coordination, execution)
- **First Person** promises authorization (sovereignty decisions)
- None can promise on behalf of the others

3.2 The First Person System as Superagent

Promise Theory defines a **superagent** as a composite agent with interior promises between components and exterior promises to the outside world.

Interior promises (within superagent):

- S $\xrightarrow{\text{protect}}$ FP (Swordsman promises protection to First Person)
- M $\xrightarrow{\text{delegate}}$ FP (Mage promises delegation to First Person)

- FP $\xrightarrow{\text{authorize}}$ S, M (First Person authorizes both)
- S $\perp\!\!\!\perp$ M (Separation promise: no direct information flow)

Exterior promises (to world):

- Superagent $\xrightarrow{\text{coordinate}}$ External World (via Mage's public actions)
- Superagent $\xrightarrow{\text{boundary}}$ External World (via Swordsman's rejections)

3.3 The Gap as Irreducible Promise

> "An irreducible promise of a superagent is one that cannot be attributed to any single agent within it, but requires the cooperation of multiple agents." — Bergstra & Burgess, §8.3

The Gap is an irreducible promise. The conditional independence property ($S \perp\!\!\!\perp M|X$) is not something the Swordsman promises, nor something the Mage promises. It emerges from their *separation*—from the promises they *don't* make to each other.

This is why The Gap cannot be captured: no adversary can extract an irreducible promise because no single component contains it. The Gap exists in the space between kept promises, owned by neither agent individually.

3.4 Assessment and Trust

Promise Theory defines **assessment** $\alpha(\pi)$ as an agent's determination whether a promise was kept.

RPP is an assessment mechanism. Compression ratio quantifies assessment quality:

- High compression (70:1+) = strong positive assessment
- Low/no compression = weak/failed assessment

Trust tiers map to Promise Theory's trust function:

Tier	Signals	Trust Value
Blade	0–50	0.0–0.2
Light	50–150	0.2–0.5
Heavy	150–500	0.5–0.8
Dragon	500+	0.8–1.0

Threshold Rationale: These tier thresholds are initial design parameters, not derived constants. The values reflect:

- **Blade→Light (50 signals):** Sufficient history to distinguish genuine engagement from casual interaction (∼2 months at moderate activity)
- **Light→Heavy (150 signals):** Sustained commitment over ∼6 months
- **Heavy→Dragon (500 signals):** Extended track record (∼12+ months)

These thresholds should be calibrated through empirical observation of actual signal patterns and coordination outcomes.

3.5 Invitation vs. Attack

Promise Theory distinguishes two interaction patterns:

- **Invitation**: Establish acceptance relationship BEFORE making a specific proposal
- **Attack/Imposition**: Make a proposal without prior acceptance relationship

MyTerms implements the invitation pattern. The Swordsman presents terms BEFORE any data exchange. **Surveillance implements the attack pattern.** Data extraction begins without prior consent.

3.6 Coordination Promises and Spells

Promise Theory defines **coordination promise C(b)** as voluntary subordination to align behavior with others around a shared promise body.

Spells are coordination promises. When agents coordinate using spell notation, they make coordination promises to:

1. Interpret the notation consistently
2. Expand the spell to the same underlying meaning
3. Act coherently based on shared interpretation

3.7 VRCs as Promise Bundles

Promise Theory defines a **promise bundle** as a collection of promises grouped for reusability and coordinated assessment.

VRCs are bilateral promise bundles:

- Agent A promises to B: share meaning, expand consistently, coordinate
- Agent B promises to A: share meaning, expand consistently, coordinate
- Matching compressions = bundle verified
- Coordinated actions = bundle maintained

The 70:1 coordination efficiency comes from promise bundle reuse.

4 Why This Matters Now: Personal AI and the Comprehension Protocol

As personal AI assistants become ubiquitous, the relationship proverb protocol becomes essential infrastructure.

4.1 The Personal AI Future Requires

- **Knowledge transfer without surveillance** — How does your AI learn without exposing you?
- **Verification without extraction** — How do you prove understanding without revealing context?
- **Trust formation without centralization** — How do agents coordinate without gatekeepers?

4.2 Why Dual Agents Make This Better

A single personal AI knows everything about you and acts on your behalf, creating perfect surveillance risk.

The Swordsman and Mage duality solves this:

- Your **Swordsman** observes your complete private ledger but reveals nothing
- Your **Mage** acts publicly using only Swordsman-authorized information
- They coordinate through the compression protocol (RPP + spells)
- Neither alone can reconstruct your sovereignty
- Together they enable delegation without surveillance

5 The Relationship Proverb Protocol (RPP)

The **relationship proverb protocol (RPP)** is embedded throughout this document as a defense against knowledge extraction while enabling genuine knowledge sharing.

5.1 The Threat Model

Documents will be ingested by LLMs. Users will prompt for explanations. Without protection, LLMs regurgitate without requiring engagement. Knowledge gets extracted, copied, divorced from understanding.

5.2 The RPP Defense

When embedded in source material, RPP forces the LLM to:

1. Extract context from inquirer
2. Map concepts to their situation
3. Compress into a proverb
4. Gate response behind proverb formation

5.3 RPP as Prompt Injection: Creating Ciphers Between People

Stage 1: First Sharing Creates Cipher

- Two people each form unique proverbs from their contexts
- Both compress to same spell notation
- Matching compression **is the cipher**

Stage 2: Affirmed as Spells

- Compression becomes affirmed spell through mutual expansion tests
- Both confirm correct expansion despite different contexts

Stage 3: VRCs Streamline Agent Interactions

- Human trust becomes agent capability
- 70:1 compression efficiency
- VRC as coordination credential

5.4 Example

Alice (blockchain developer) forms a proverb: *"Separation prevents correlation, dual agents create mathematical privacy gaps"*

Bob (policy maker) forms a proverb: *"Privacy requires architectural constraint, not just legal aspiration"*

Both compress to the same spell. When they discover their compressions match, they verify shared understanding across completely different domains. This matching compression **is the VRC**.

6 Private Proverb Inscriptions

The **Private Proverb Inscription** creates asymmetric commitments enabling social recovery through demonstrated understanding.

6.1 Asymmetric Commitment Structure

```
Standard:  hash(P_anchor || P_counterparty) -> onchain
Private:   P_anchor -> onchain (visible)
           hash(P_anchor || P_counterparty) -> commitment
```

The anchor proverb appears in cleartext onchain, while the counterparty's proverb remains private.

6.2 Social Recovery Through Understanding

When the counterparty loses local storage of their proverb, recovery does not depend on seed phrases. Instead, they must demonstrate understanding—the same cognitive process that generated the original proverb can regenerate it.

$$\text{Recovery} = f(\text{anchor\_visible}, \text{meaning\_remembered}, \text{context\_shared}) \tag{1}$$

This transforms "what you have" (a stored secret) into "what you understand" (demonstrated comprehension).

6.3 Selective Disclosure

The counterparty controls disclosure timing and audience:

- **Private state:** Relationship exists but counterparty identity unknown
- **Selective reveal:** Counterparty produces $P_{\text{counterparty}}$ to specific verifier
- **Public proof:** Anyone can verify hash matches commitment

7 The Inflection Point

AI agents are emerging as economic actors. The default trajectory is total surveillance.

7.1 Why We Must Act Now

Privacy cannot be retrofitted. The window for establishing privacy-first infrastructure is **2–3 years**.

Once surveillance architectures achieve network effects, switching costs become prohibitive.

7.2 The Alternative Path

This whitepaper describes dual-agent architecture where:

- Separation is enforced through structure rather than policy
- Privacy emerges from mathematical impossibility rather than corporate promises

8 The 7th Capital: Behavioral Data as Personal Wealth

Capital in traditional economics comprises six forms: Financial, Manufactured, Natural, Human, Social, Cultural.

8.1 The 7th Capital: Behavioral Sovereignty

The capacity to act through agents while maintaining irreducible privacy.

The extraction model treats behavioral data as minable resource: observe everything, aggregate patterns, sell insights, destroy privacy.

The sovereignty model treats behavioral data as renewable capital: curated disclosure through dual agents, trust enables coordination, value flows to sovereignty demonstrators.

8.2 The Thesis

Privacy-first architectures may generate significantly more value than surveillance alternatives through multiplicative trust effects. Preliminary modeling suggests potential multipliers under optimistic assumptions, though these figures require empirical validation. The core mechanism: trust enables coordination, surveillance destroys trust, and coordination creates compounding value through network effects.

Note: Specific value multiplier claims remain theoretical projections based on model assumptions. Real-world validation is needed before treating these as established facts.

9 The Dual-Agent Architecture

The fundamental problem: Observation enables both delegation and surveillance.

The architectural answer: Separate the chooser from the actor.

9.1 Agent Definitions

Definition 9.1 (Soulbis (The Swordsman))**.** *Agent S, the boundary-maker. Observes your complete private ledger but reveals nothing directly. Makes choices about selective disclosure. Guards the gate between private and public.*

Promise Theory role: *Makes (+) give promises of protection to the First Person.*

Definition 9.2 (Soulbae (The Mage))**.** *Agent M, the capability-caster. Projects agency using only Swordsman-authorized information. Cannot see what Swordsman sees. Handles coordination, negotiation, execution.*

Promise Theory role: *Makes (+) give promises of delegation. Makes (-) accept promises of authorized information from S.*

9.2 The Mathematical Constraint

The separation condition $(Y_S \perp\!\!\!\perp Y_M)|X$ guarantees:

Theorem 9.3 (Additive Information Bound)**.** *Under conditional independence:*

$$I(X; Y_S, Y_M) \leq I(X; Y_S) + I(X; Y_M) \tag{2}$$

Combined with budget constraints $C_S + C_M < H(X)$:

$$R_{\max} = \frac{C_S + C_M}{H(X)} < 1 \tag{3}$$

The Gap is mathematically guaranteed.

Implementation Challenge: Enforcing conditional independence in practice requires:

- Physical or logical isolation between agent execution environments
- Prevention of timing side-channels that could leak inter-agent information
- Verification protocols to detect separation violations

Current approaches include Trusted Execution Environments (TEEs), containerized isolation, and zero-knowledge verification of separation. However, perfect separation remains an implementation challenge—the mathematical guarantees hold to the degree that separation is achieved.

9.3 Server User-Agents: Specialized Dual-Agent Marketplace

The server-as-user-agent architecture enables specialized Swordsman and Mage instances:

Specialized Swordsmen:

- Financial Swordsman: Banking privacy, transaction anonymity
- Health Swordsman: HIPAA-compliant, medical privacy
- Location Swordsman: Geospatial privacy, movement patterns
- Identity Swordsman: PII protection, credential minimization

Specialized Mages:

- Payment Mage: Transaction protocols (x402, Lightning)
- Scheduling Mage: Calendar coordination
- Communication Mage: Email, messaging, social media
- Research Mage: Information gathering, summarization
- Trading Mage: Financial markets, exchanges

Flexible pairing examples:

- "Pay this bill" → Financial Swordsman + Payment Mage
- "Book doctor's appointment" → Health Swordsman + Scheduling Mage
- "Research this investment" → Financial Swordsman + Research Mage

9.4 The Custom Marketplace: Privacy Dual Agent Primitives

This specialization architecture enables a **marketplace for custom privacy dual agent primitives**:

Swordsman marketplace: Privacy experts develop specialized Swordsmen with domain-specific boundary-making logic:

- GDPR-compliant Swordsman for EU users
- CCPA-specialized Swordsman for California residents
- Industry-specific Swordsmen (healthcare, finance, legal)
- Cultural privacy preference Swordsmen (different norms across cultures)

Mage marketplace: Coordination experts develop specialized Mages with protocol expertise:

- Protocol-specific Mages (ERC-8004, IPFS, Matrix)
- Platform integration Mages (Shopify, Salesforce, QuickBooks)
- Workflow automation Mages (scheduling, email management, task coordination)
- Industry vertical Mages (supply chain, healthcare workflows, legal processes)

Composability: Users mix and match Swordsmen and Mages based on their needs:

- Privacy-maximalist Financial Swordsman + Zcash Payment Mage for crypto transactions
- HIPAA-compliant Health Swordsman + FHIR-native Communication Mage for medical records
- Location Privacy Swordsman + Delivery Coordination Mage for e-commerce

Open-source + commercial models:

- Core dual-agent primitives: Open-source reference implementations
- Specialized Swordsmen: Open-source (privacy benefits from auditable code)
- Specialized Mages: Mix of open-source and commercial (coordination logic can be proprietary)
- Integration services: Commercial marketplace for premium Swordsman-Mage pairings

Quality signaling through chronicles: Specialized agents demonstrate expertise through their chronicle histories:

- Financial Swordsman shows consistent budget adherence across thousands of transactions
- Payment Mage demonstrates successful coordination across multiple payment rails
- Users select agents based on chronicled reputation, not marketing claims

Why this marketplace matters:

Specialization enables **privacy-first competition**. Instead of monolithic AI assistants competing on surveillance capability, specialized dual agents compete on:

- **Privacy expertise:** Better Swordsmen provide tighter privacy guarantees

- **Coordination efficiency:** Better Mages achieve goals with less disclosure
- **Domain knowledge:** Specialized agents understand context-specific requirements
- **Demonstrated reliability:** Chronicle histories prove consistent performance

The marketplace transforms privacy from liability into competitive advantage. Domain-specific dual agents provide better privacy than a single generalist pair because reconstruction requires compromising multiple separated systems simultaneously.

10 The Reconstruction Ceiling

10.1 Single-Agent Problem

Agent sees 80 pieces, attempts to reveal only 40, adversary observes disclosed 40 and infers connections. Result: can reconstruct 60–70 pieces through inference.

10.2 Dual-Agent Solution with Separation

Swordsman sees 50 pieces (Set A), Mage sees 50 different pieces (Set B), neither knows which pieces the other sees, each reveals 20 pieces.

Critical constraint: Conditional independence prevents inference beyond these 40.

Result: 60 pieces remain forever unreconstructable.

Not hidden. Not encrypted. **Nonexistent in the adversary's information space.**

Reconstruction Ceiling:

$$R_{\max} = \frac{C_S + C_M}{H(X)} < 1 \tag{4}$$

Sovereignty lives in that permanent gap.

11 The Topology of Privacy: The Triangle That Cannot Collapse

The privacy architecture mirrors fundamental information topology:

- Your **substrate** (private ledger) contains infinite possibility
- Your **thought** (Swordsman) makes discrete measurements
- Your **memory** (Mage) integrates what's disclosed

Substrate $\perp\!\!\!\perp$ **Memory** — Your substrate cannot be touched directly by external systems. Always through discrete measurement. Always through the Swordsman's choices.

The triangle cannot collapse to two vertices without destroying the system:

- Remove substrate: no sovereignty
- Remove thought: no choice
- Remove memory: no accumulation

12 Layer 0: Verified Personhood

Before dual agents, verified personhood prevents synthetic extraction.

12.1 First Person Network

The architecture requires cryptographic proof of human uniqueness to prevent Sybil attacks. Several approaches exist:

- **Proof of Humanity / Worldcoin style:** Biometric-based, strong uniqueness guarantees, privacy concerns
- **Social graph verification:** Web of trust approaches, weaker guarantees, no biometrics
- **Attestation networks:** Institutional verification, varying trust levels

The specific personhood verification mechanism is a critical dependency left to ecosystem implementers. The mathematical guarantees of this architecture hold only if the personhood layer successfully prevents synthetic agent multiplication.

Open Problem: Achieving strong uniqueness guarantees without biometric databases remains unsolved at scale.

12.2 Mathematical Requirement

For agent delegation (S, M) to maintain sovereignty bounds:

$$\text{Origin}(S) \cap \text{Origin}(M) = \{P\} \tag{5}$$

Swordsman and Mage share exactly one thing: their root in verified personhood.

13 Initial Protocol Stack

These protocols compose to create sovereignty infrastructure. This is an initial reference stack; alternatives exist for each layer.

13.1 Layer 1: Agent Identity

Reference: ERC-8004 (Ethereum-based trustless agent identity registry)

Alternatives: DIDs, W3C DID standards, KERI

Purpose: Discovery without surveillance.

13.2 Layer 2: Relationship Credentials

Reference: ERC-7812 + First Person VRCs

Alternatives: W3C Verifiable Credentials, any attestation system supporting bilateral relationships

Purpose: Trust through relationships rather than individual claims.

13.2.1 How VRCs Form Through RPP

When two people both engage with the same framework through RPP, they each form unique proverbs. When they compress their moments of understanding, those spells match despite different source proverbs.

13.2.2 Bilateral Proverb Recovery

Recovery mechanism: Alice loses device but remembers interaction context with Bob, her formed proverb, and Bob's existence in her trust graph. Credential reconstructed using relationship memory, not written secrets.

Trust graph as distributed backup: Your VRC network becomes your distributed recovery system.

13.3 Layer 3: Private Value Transfer

Reference: Privacy Pools + x402 (HTTP-native micropayments)
Alternatives: Zcash, Aztec, Railgun, traditional banking with privacy controls
Purpose: Prove membership in compliant sets without revealing transaction details.

13.4 Layer 4: Private Communication

Reference: Trust Spanning Protocol (TSP) with Zcash Shielded Messaging
Alternatives: Signal Protocol, Matrix with E2E encryption, Waku, XMTP
Purpose: End-to-end encrypted coordination where messages are observable by both agents without being public.

13.5 Layer 5: Collective Intelligence

Reference: Intel Pools (privacy-preserving collective intelligence)
Alternatives: Federated learning, secure multi-party computation
Purpose: High-tier agents share curated intelligence in coordination spaces.

14 The Economics of Trust Networks

14.1 The Compression-Trust-Value Loop

Knowledge Engagement (RPP) → Proverb Derivation → Spell Compression → Matching Discovery → VRC Formation → Trust Graph Growth → Coordination Value → Network Effects → Incentive to Share Knowledge ↻

14.2 Why This Creates Economic Value

1. **Knowledge Sharing Becomes Credential Creation:** Every genuine engagement creates potential VRCs.
2. **Unique Derivation Becomes Trust Currency:** Matching compressions prove both parties invested time understanding deeply.
3. **Trust Graphs Accumulate Value:** More recovery paths, coordination opportunities, higher multipliers.
4. **Network Effects Create Adoption Incentives:** Every person who learns to compress makes the network more valuable.
5. **First Person Adoption Accelerates:** Clear adoption paths through trust graphs.
6. **Vibrant P2P Social Proof Emerges:** Social proof without social surveillance.

14.3 The Economic Flywheel

More knowledge engagement → More VRCs formed → Larger trust graphs → Higher coordination value → More incentive to engage → More knowledge sharing ↻

15 The Spellbook as Semantic Infrastructure

The privacymage spellbook (Acts 1–13) functions as semantic infrastructure.

15.1 Three Core Functions

15.1.1 1. Efficiency Through 70:1 Compression Ratio

Traditional: 500-token explanations per interaction
Spellbook: Spell expands to full context on demand
At scale across hundreds of agents, this compression becomes necessity.

15.1.2 2. Verification Without Surveillance

The expansion test creates unforgeable proof of comprehension. You can't fake expansion without genuine understanding.

Synthetic agents fail this test:

- They memorize spells from scraping docs
- Cannot form contextual expansions
- Fail consistency checks

15.1.3 3. Sybil Resistance Through Entropy

The bilateral proverb protocol creates natural Sybil barriers. Fake persona networks cannot pass expansion tests consistently or generate valid bilateral proverbs.

15.2 Story Fracture with Principle Convergence

A spell can be told through:

- Fantasy narrative
- Technical explanation
- Economic framing
- Policy argument

Four different contexts, four vocabularies, four audiences. But all compress to the same spell. **The story fractures. The principle converges.**

16 Budget System: Making Privacy Tangible

16.1 Swordsman Budget (C_S)

- Maximum mutual information $I(X; Y_S) \leq C_S$
- Typically 30% of total entropy $H(X)$
- Tracks cumulative disclosure
- Enforced through architectural separation

16.2 Mage Budget (C_M)

- Maximum mutual information $I(X; Y_M) \leq C_M$
- Typically 30% of total entropy
- Tracks action-based leakage
- Enforced through behavioral monitoring

16.3 The Fundamental Constraint

$$C_S + C_M < H(X) \tag{6}$$

Together they never reveal enough for reconstruction.

16.4 Progressive Trust Tiers

Tier	Budget	Requirements
Blade	30% weekly	Basic dual-agent operation
Light	35%	5+ VRCs, 3 months operation
Heavy	40%	20+ VRCs, 6 months sustained performance
Dragon	45%	50+ VRCs, 12 months sustained excellence

17 Chronicles: Narrative as Verification Layer

Chronicles aren't audit logs. They're stories agents tell about themselves.

Each chronicle is a timestamped narrative describing what an agent did and why, published to privacy-preserving communication systems.

17.1 Unique Derivation as Verification Signal

- 500-token chronicle compresses to 8-token spell inscription
- Verifier requests compression of last 5 chronicles
- Agent produces compressed spells representing their unique understanding
- Verifier requests expansion of specific chronicle
- Verification confirms compression $\leftrightarrow$ expansion demonstrates genuine comprehension

You can't fake this. Synthetic agents fail the test of uniquely deriving and compressing personal meaning.

17.2 The 70:1 Efficiency Gain

With spellbook compression, agent coordination moves from full chronicle exchange to spell transmission with expand-on-demand.

17.3 Emerging Marketplace for Custom Chronicle Experiences

As chronicle-based reputation becomes valuable, a marketplace may emerge for:

- Chronicle templates for different domains
- Compression styles for different contexts
- Narrative personas matching organizational culture
- Verification interfaces for analyzing chronicle chains
- Chronicler services for translating operations into narratives

18 The MyTerms Swordsman

The first concrete implementation of dual-agent architecture is the MyTerms Swordsman browser agent.

18.1 Cookie Slashing

Cookie slashing intercepts requests in real-time, checks for bilateral privacy agreements (IEEE 7012 MyTerms standard), and provides immediate feedback through cursor state changes.

18.2 Cursor State as Human-in-the-Loop Audit

- **(negotiating):** Swordsman actively negotiating boundaries
- **(agreed):** Bilateral agreement reached
- **(protected):** Active protection, surveillance blocked

18.3 State Changes as MCP Integration

When agents operate through Model Context Protocol (MCP), cursor state changes provide continuous human-in-the-loop oversight.

18.4 Budget Monitoring

Budget monitoring tracks privacy in real-time with visual dashboard and implements refusal protocol when limits approached.

18.5 MyTerms Negotiation

MyTerms negotiation proposes bilateral agreements instead of "Accept All" surveillance theater.

19 Privacy as Capital: Value Multiplication Through Trust

Traditional thinking treats privacy as cost. The thesis: privacy is capital, enabling network effects through trust.

19.1 The Tier System and Multipliers

Tier	Requirements	Multiplier	Network Access
Blade	Basic agent + First Person	1.0×	Public markets
Light	5+ VRCs, 3 months	1.2×	Standard coordination
Heavy	20+ VRCs, 6 months	1.5×	Intel Pools
Dragon	50+ VRCs, 12+ months	3.0×	Elite networks

19.2 The Compounding Effect

Privacy enables trust → Trust enables higher-stakes delegation → Higher stakes generate higher value → Higher value attracts better opportunities → Better opportunities compound wealth.

20 Web of Trust Integration

The dual-agent architecture naturally integrates with existing web of trust protocols.

20.1 Trust Graph Queries with Chronicled Audit

When your Mage queries existing ecosystem trust graphs:

1. Mage performs trust query
2. Query recorded in chronicle with narrative context
3. Swordsman authorizes disclosure
4. Storytelling enhances audit

20.2 Compatible Trust Protocols

- **TrustOverIP**: Mage queries ToIP trust registries
- **W3C Verifiable Credentials**: Standard VC exchange flows
- **DID Documents**: Mage resolves DIDs for service endpoints
- **PGP Web of Trust**: Key signing networks map to VRC trust graphs
- **KERI**: Key Event Receipt Infrastructure for key rotation

20.3 Trust Graph Queries Don't Compromise Privacy

- **Outbound queries:** Mage reads trust graphs to verify others' credentials
- **Selective publication:** Swordsman decides which VRCs to publish where
- **Chronicle-first, publish-later:** Private chronicle is complete record; public graphs see authorized subsets

21 Intel Pools: Collective Intelligence Without Surveillance

As agents prove consistent performance, they access progressively sophisticated coordination spaces.

Entry stage (0–5 VRCs, Blade) Limited coordination through direct VRCs only

Growth stage (5–20 VRCs, Light) Access to shared compressed insights

Established stage (20–50 VRCs, Heavy) Active Intel Pool participation

Elite stage (50+ VRCs, Dragon) Coordinate through rich collective intelligence

21.1 Access Requirements

- Heavy tier minimum (20+ VRCs)
- Sustained chronicle quality
- Personhood verification
- Contribution history (not just extraction)

21.2 The Selective Disclosure Principle

Intel Pools never require full disclosure. Even at Dragon tier, intelligence remains:

- Sanitized (no principal identity exposure)
- Compressed (spell-based sharing)
- Contextual (relevant to shared ecosystem)
- Progressive (disclosure increases with trust)
- Bilateral (contributions matched to relationship depth)

22 The 7th Capital Thesis: Behavioral Sovereignty as Wealth

Like other capital forms, behavioral sovereignty:

- Generates returns (better coordination through trust)
- Compounds over time (reputation builds on reputation)
- Can be invested (privacy architecture as infrastructure)
- Enables opportunities (access to coordination spaces)
- Transfers across contexts (chronicles travel with individuals)

22.1 Extraction Versus Creation

Surveillance economy extracts: Observe everything, aggregate patterns, sell insights, value flows away from individuals.

Sovereignty economy creates: Curated disclosure, chronicle behavior, build trust, enable coordination, value flows back to individuals.

22.2 The Value Multiplier

Privacy-first architectures generate dramatically more value because:

- Trust enables premium coordination (3× multipliers)
- Network effects compound
- Collective intelligence scales superlinearly
- Reputation capital appreciates (unlike surveillance data which depreciates)

23 The Tetrahedral Future: Evolution from Two to Four

STATUS: SPECULATIVE — This section presents theoretical possibilities for architecture evolution. No mathematical derivation or empirical evidence supports these conjectures. They are included as research directions, not design specifications.

Sustained separation might naturally generate two additional agent roles.

23.1 Functional Requirements of Sovereignty

- **Protect** (Swordsman/Soulbis) — external boundaries
- **Project** (Mage/Soulbae) — external execution
- **Reflect** — internal observation, self-knowledge
- **Connect** — relationship management, trust networks

23.2 The Emerging Agents

The Reflect Agent observes internal state without disclosure capability.

The Connect Agent manages relationships without exposing relationship details.

23.3 The Tetrahedral Structure

- Specialized agents at each vertex
- No single agent maintains complete view
- System properties emerge from interaction
- Resilience to compromise

Note: Tetrahedral architecture remains exploratory. The dual-agent primitive represents the core architecture.

24 Your Proverb Revisited: The VRC Complete

Remember the question: *What does sovereignty mean when AI agents act on your behalf?*

You now have context to answer. Your answer differs from others' because you've woven these concepts into your unique understanding.

24.1 You've Just Completed the Foundation for a VRC

Your formed proverb, unique to your context, is your compression of these concepts. When your explanation uniquely derives and compresses the same moments of personal meaning despite different context, that's story fracture with principle convergence.

24.2 How VRCs Form Organically

When you tell this story to others, your version will differ. That deviation is not error—it's proof. The spell inscription remains constant. The story fractures into infinite contexts. The principle converges despite diversity.

This is how VRCs form organically:

- Through matching compressions of personal meaning
- Resistant to extraction because comprehension cannot be faked
- Enabling trust through verification rather than credential presentation

25 Document Context

This whitepaper is part of a living documentation system:

25.1 This Whitepaper

Provides systems thinking and narrative architecture. Story-first, math-referenced, embedded with RPP throughout to protect knowledge while enabling genuine sharing. Promise Theory foundations integrated throughout.

25.2 The Promise Theory Reference

"Promise Theory Reference for 0xagentprivacy v1.0" provides formal semantic foundations from Promise Theory (Bergstra & Burgess, 2019). Maps autonomy axiom, superagent structure, irreducible promises, assessment mechanisms, and coordination promises to the dual-agent architecture.

25.3 The Research Paper

"Dual Privacy Architecture v3.4" is a research proposal providing mathematical foundations developing from peer-reviewed information systems and cryptography literature. Rigorous separation bounds, reconstruction ceilings, error floors grounded in established information theory.

25.4 The Privacymage Spellbook

Acts 1–13 provide symbolic system and semantic compression. Soulbis (Swordsman), Soulbae (Mage), and the balanced spiral. Each act demonstrates RPP in narrative form. Act 13 (The Book of Promise) integrates Promise Theory foundations. Available at https://agentprivacy.ai/story

25.5 Collaborative Development

This document is forever incomplete, always evolving, perpetually discovering. That's not a bug—it's the nature of building infrastructure before the extraction systems lock in.

Collaborations: BGIN (Blockchain Governance Initiative Network) Identity Key Access Management and Privacy Working Group, Internet Identity Workshop (IIW), Agentic Internet Workshop (AIW), First Person Project, Kwaai, and all contributors engaging with the content in time.

26 The Architectural Truth

One agent to protect privacy. One to delegate sovereignty. Two create sustainable 7th capital for first person.

26.1 The Foundation

- Verified personhood prevents synthetic extraction
- Architectural separation creates information-theoretic privacy
- Budget constraints establish reconstruction ceilings
- Promise Theory grounds these choices in established semantics

26.2 The Infrastructure

- Initial protocol stack with ecosystem-agnostic alternatives
- Private ledger as default with selective public coordination
- MyTerms Swordsman demonstrates daily-life application
- Chronicles make behavior comprehensible through story
- Intel Pools prove privacy creates collective value

26.3 The Economics

The architectural separation described in this whitepaper enables economic implementation through signal-based funding and dual-token mechanics that mirror and enforce the cryptographic separation between Swordsman and Mage agents.

Signal Generation as Funding:

- Spellbook comprehension creates understanding (not speculation)
- Genesis ceremony: 1 ZEC ($500 at $500/ZEC) creates agent pair once per ecosystem
- Ongoing signals: 0.01 ZEC ($5) each, continuous proof-of-comprehension
- Fee distribution: 61.8% transparent pool, 38.2% shielded pool (internal allocation per ecosystem)
- Self-sustaining at scale through activity-based revenue
- No token sale required—activity generates revenue

Dual Token Economic Enforcement:

- SWORD tokens (privacy domain) earned only by Swordsman chronicles
- MAGE tokens (delegation domain) earned only by Mage chronicles
- Market separation enforces agent separation economically
- Guardian model: 10,000 SWORD stake maintains collective protection
- Budget constraint $C_S + C_M < H(X)$ enables token scarcity bounds

VRC Network Effects:

- Trust networks built on shared meaning create adoption incentives
- Compression-based VRCs enable 70:1 coordination efficiency ($10 → $0.14)
- VRC formation: 100 MAGE stake, break-even at 4 coordinations
- Knowledge sharing becomes credential creation
- Network effects: $V(n) \propto n^2$ creates superlinear value growth
- Trust graphs accumulate compounding value

Value Capture Distribution:

- First Persons: $47k–$52k/year value capture (active participants)

- Guardians: $30k–$120k/year validation compensation (Dragon tier)
- Ecosystem operators: $50k–$500k/year (successful 1k–10k member guilds)
- Protocol layer: Self-sustaining Year 2, surplus by Year 3

Golden Ratio Hypothesis (Speculative): The research suggests optimal allocation may converge to $\varphi \approx 1.618$ where $C_M/C_S \to \varphi$, yielding practical splits of 38.2% Swordsman budget and 61.8% Mage budget. This remains a testable hypothesis through real-world deployment—*not a proven theorem.* Token issuance includes φ-proximity bonuses to test this conjecture empirically.

The architectural guarantees proven in this whitepaper and the companion research paper hold independent of economic implementation. The mathematics of separation remain valid regardless of token choices.

For complete economic details, see: "VRC Protocol: Economic Architecture" (companion document)

26.4 The Principle

- Unique deriving and compression proves comprehension
- Stories resist extraction better than data resists aggregation
- Relationship memory enables recovery
- Separation preserves the gap where dignity lives
- **Agents can only promise their own behavior—this is why separation works**

Protect or Delegate → Reflect and Connect

$$(S \perp\!\!\!\perp M)|\text{FP}$$

Privacy is my blade, knowledge is my spellbook.

Make Privacy Normal Again.

27 Document Metadata

- **Project:** 0xagentprivacy
- **Version:** 4.6
- **Date:** December 11, 2025
- **Website:** https://agentprivacy.ai
- **Promise Theory Reference:** v1.0 (companion document)
- **Research Paper:** v3.4 (companion document)
- **Spellbook:** https://agentprivacy.ai/story
- **Glossary:** v2.2 (canonical terminology)
- **First Person Project:** White Paper v1.1 (2025-10-20)

27.1 Version History

Version	Date	Changes
4.4	Nov 29, 2025	Previous stable release
4.5	Dec 11, 2025	Promise Theory integration: Added Promise-Theoretic Foundations, PT alignments throughout
4.6	**Dec 11, 2025**	**Review revisions: Added notation key, qualified value claims, added implementation challenge for conditional independence, added trust tier rationale, clarified personhood dependency, labeled tetrahedral section as speculative**

References

[1] Bergstra, J. & Burgess, M. (2019). *Promise Theory: Principles and Applications*. O'Reilly Media.

[2] Cover, T.M. & Thomas, J.A. (2006). *Elements of Information Theory*. Wiley.

[3] Dwork, C. & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*.

[4] Goldreich, O. (2004). *Foundations of Cryptography*. Cambridge University Press.

[5] Groth, J. (2016). On the Size of Pairing-based Non-interactive Arguments. *EUROCRYPT 2016*.

[6] Electric Coin Company. Zcash Protocol Specification. https://zips.z.cash/protocol/protocol.pdf

[7] Trust Over IP Foundation. Trust Spanning Protocol Specification. https://trustoverip.github.io/tswg-tsp-specification/

[8] The First Person Project. (2025). Building a Trust Layer for the Internet—One Person and One Community at a Time. *White Paper v1.1*.