Full Knowledge Proofs: Probabilistic Consensus for Machine Cognition

Abstract

The autonomous AI economy is projected to reach $30 trillion by 2030, with AI agents executing billions of micro-transactions across decentralized networks. Yet this future faces an existential threat: we cannot trust black-box AI systems with consequential decisions. Current verification approaches, particularly zero-knowledge proofs, prioritize privacy over transparency—hiding the very reasoning we need to audit.

We introduce full knowledge proofs (FKPs), a new cryptographic primitive that establishes probabilistic consensus for machine cognition. Where ZKPs verify computation while hiding it, FKPs verify reasoning while revealing it through interactive reasoning cycles. Our protocol coordinates multiple AI models in recursive interaction, forcing convergence to statistically significant consensus while generating transparent reasoning trails. Each reasoning cycle is cryptographically committed using a novel commitment tree structure, with economic security ensured through stake-based slashing.

The core innovation lies in our multi-model extraction lemma, which proves that any adversary successfully forging consensus must either reveal valid reasoning or break the discrete logarithm problem. We provide rigorous security proofs demonstrating completeness, soundness, and knowledge soundness under standard cryptographic assumptions.

Unlike ensemble methods that merely aggregate predictions, FKPs create verifiable audit trails of how conclusions emerged through interactive reasoning. Unlike explainable AI techniques that provide post-hoc justifications, FKPs provide cryptographically guaranteed reasoning transparency. And unlike traditional multi-prover interactive proofs, our protocol enables structured model interaction at scale while maintaining verifiable consistency.

1 Introduction

The emergence of autonomous AI agents managing critical infrastructure, financial systems, and healthcare decisions represents both unprecedented opportunity and existential risk. While zero-knowledge proofs [GMR85] have revolutionized verifiable computation, they address the wrong problem for autonomous AI: they prioritize privacy when we need transparency, they verify computation when we need to verify reasoning, and they provide cryptographic certainty about mathematical operations when we need statistical confidence about cognitive processes.

1.1 Trust Gap

Current AI systems operate as black boxes [LBD+89]. When an AI model makes a medical diagnosis, executes a financial trade, or generates legal analysis, we receive only the output—not the reasoning process. This creates a fundamental trust gap that prevents scaling autonomous systems to their projected $30T potential [Gartner2030].

Existing solutions fail to bridge this gap:

- **Zero-Knowledge Proofs** [Gro10,BCI+13]: Verify computation but hide reasoning
- **Explainable AI** [GVR17]: Provide post-hoc justifications without verification
- **Ensemble Methods** [Die00]: Aggregate predictions without transparency
- **Multi-Prover Proofs** [BGKW88]: Prevent communication between provers

1.2 Interactive Reasoning Cycles

We introduce full knowledge proofs, characterized by interactive reasoning cycles that enable:

1. **Transparent verification** of machine cognition through multi-model interaction
2. **Probabilistic consensus** emerging from recursive reasoning cycles
3. **Economic security** through stake-based slashing of reasoning participants
4. **Cryptographic audit trails** of reasoning formation across cycles
5. **Statistical confidence measures** for consensus quality

Our protocol transforms AI verification from "trust the output" to "verify the reasoning process," enabling the trust layer required for scalable autonomous economies through structured interactive reasoning cycles.

1.3 Proof Systems

The foundation of zero-knowledge proofs was established by Goldwasser, Micali, and Rackoff [GMR85], with subsequent work on non-interactive proofs [BFM88] and multi-prover systems [BCC88]. Pedersen commitments [Ped91] provide the binding and hiding properties essential for our construction, while -protocols [Cra97] inspired our interactive structure.

Recent advances in succinct proofs [Gro10, BCI+13, BBHR18] have enabled efficient verification of complex computations, but maintain the zero-knowledge property. Our work represents a philosophical inversion of this paradigm.

1.4 AI Safety

The AI alignment problem [Soa16, Rus19] emphasizes ensuring AI systems behave according to human values. Explainable AI techniques [GVR17, RSG16, LL17] provide interpretability but lack verification guarantees. Ensemble methods [Die00] and mixture of experts [JSS90] combine multiple models but focus on accuracy rather than transparency.

Recent work on verifiable AI [KLV+22] has explored formal verification of neural networks, but typically focuses on specific properties rather than general reasoning verification.

1.5 Blockchain

Oracle networks like Chainlink [EZ19] bring external data on-chain, while decentralized AI platforms [Fet19, OC21, Bit21] focus on model training and inference. Our work complements these by providing a verification layer for AI reasoning processes through interactive reasoning cycles.

1.6 Cryptographic Foundations

Definition 1.1 (Discrete Logarithm Problem)**.** Let $\mathbb{G}$ be a cyclic group of prime order p with generator g. The Discrete Logarithm Problem (DLP) is hard in $\mathbb{G}$ if for all probabilistic polynomial-time adversaries $\mathcal{A}$:

$$\Pr\left[\mathcal{A}(g, g^a) = a\right] \leq \mathsf{negl}(\lambda)$$

where $a \xleftarrow{\$} \mathbb{Z}_p$ and λ is the security parameter.

Definition 1.2 (Pedersen Commitment)**.** The Pedersen commitment scheme consists of:

$$\begin{aligned}\mathsf{Setup}(1^\lambda) &: \text{Output } pp = (\mathbb{G}, p, g, h) \\ \mathsf{Commit}(m, r) &= g^m h^r \mod p \\ \mathsf{Open}(c, m, r) &: \text{Verify } c = g^m h^r\end{aligned}$$

The scheme is perfectly hiding and computationally binding under DLP.

1.7 Machine Cognition

Definition 1.3 (Interactive Reasoning Cycle)**.** A reasoning cycle consists of multiple AI models independently processing inputs at scale, then sharing and refining their reasoning based on peer outputs. Each cycle produces:

- Updated outputs based on peer reasoning
- Refined confidence scores
- Cryptographic commitments to reasoning traces

Definition 1.4 (Probabilistic Consensus)**.** A sequence of reasoning cycles reaches probabilistic consensus when:

$$\lim_{r \to \infty} \max_{i,j} |y_i^{(r)} - y_j^{(r)}| < \delta$$

with statistical confidence $C > 1 - \epsilon$ for some convergence threshold δ and error bound ϵ.

2 Construction

2.1 Protocol Definition

Definition 2.1 (Interactive Full Knowledge Proof)**.** An Interactive Full Knowledge Proof system for relation R consists of:

$$\mathsf{IFKP} = (\mathsf{Setup}, \langle P, V, M_1, \ldots, M_n \rangle, \mathsf{Verify}, \mathsf{Extract})$$

where:

- $pp \leftarrow \mathsf{Setup}(1^\lambda)$: Output public parameters
- $\pi \leftarrow \langle P, V, M_1, \ldots, M_n \rangle(Q)$: Interactive proof generation through reasoning cycles
- $\{0, 1\} \leftarrow \mathsf{Verify}(pp, Q, \pi)$: Proof verification
- $w \leftarrow \mathsf{Extract}(pp, \pi)$: Witness extraction

2.2 Core Protocol

Protocol 2.1 (Full Knowledge Proof Protocol with Interactive Reasoning Cycles)**. Input:** Query Q, models $\mathcal{M} = \{M_1, \ldots, M_n\}$, max reasoning cycles R

Output: Proof $\pi = (CT, y^*, C^*, \sigma)$

1. **Initialization:**
   - Initialize commitment tree CT
   - For each model $M_i \in \mathcal{M}$:
     $$y_i^{(1)}, r_i^{(1)} \leftarrow M_i(Q)$$
     $$c_i^{(1)} \leftarrow \mathsf{com}_i^{(1)} = g^{H(y_i^{(1)})} h^{H(r_i^{(1)})}$$
   - Store $CT[1][i] = (c_i^{(1)}, y_i^{(1)}, r_i^{(1)})$
2. **Interactive Reasoning Cycles:** For cycle $r = 1$ to R:
   - Compute median: $\mu^{(r)} = \mathsf{median}(\{y_i^{(r)}\})$
   - Compute MAD: $\sigma^{(r)} = \mathsf{median}(|y_i^{(r)} - \mu^{(r)}|)$
   - Detect outliers: $O^{(r)} = \{i : |y_i^{(r)} - \mu^{(r)}| > k \cdot \sigma^{(r)}\}$
   - If $|O^{(r)}| = 0$, break (probabilistic consensus achieved)
   - For each $M_i \notin O^{(r)}$:
     $$y_i^{(r+1)}, r_i^{(r+1)} \leftarrow M_i(Q, \{y_j^{(r)}, r_j^{(r)}\}_{j \neq i})$$
     $$c_i^{(r+1)} \leftarrow g^{H(y_i^{(r+1)})} h^{H(r_i^{(r+1)})}$$
   - Store reasoning cycle results: $CT[r{+}1][i] = (c_i^{(r+1)}, y_i^{(r+1)}, r_i^{(r+1)})$
3. **Finalization:**
   - $y^* = \mu^{(r)}$ (final probabilistic consensus)
   - $C^* = 1 - \frac{|O^{(r)}|}{n} \cdot (1 - \min\{c_i^{(r)} : i \notin O^{(r)}\})$
   - $\sigma = \mathsf{Sign}(CT, y^*, C^*)$
   - Output $\pi = (CT, y^*, C^*, \sigma)$

2.3 Commitment Tree Structure

The commitment tree CT provides cryptographic binding of rounds:

$$CT = \begin{bmatrix} c_1^{(1)} & c_2^{(1)} & \cdots & c_n^{(1)} \\ c_1^{(2)} & c_2^{(2)} & \cdots & c_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ c_1^{(R)} & c_2^{(R)} & \cdots & c_n^{(R)} \end{bmatrix}$$

Each commitment $c_i^{(r)} = g^{H(y_i^{(r)})} h^{H(r_i^{(r)})}$ binds both the output and reasoning at each round, creating an immutable audit trail of the entire cognitive process.

2.4 Security Properties

Theorem 2.1 (Completeness)**.** For all $(Q, w) \in R$ and honest protocol execution:

$$\Pr[\mathsf{Verify}(pp, Q, \langle P, V, M \rangle(Q)) = 1] = 1$$

Proof. Honest execution produces valid commitment openings and satisfies all verification criteria by construction. The interactive reasoning cycles naturally converge to probabilistic consensus when models reason honestly about the same problem. □

Theorem 2.2 (Soundness)**.** If DLP is hard in $\mathbb{G}$ and at least $\lceil n/2 \rceil$ models are honest, then for all PPT adversaries $\mathcal{A}$:

$$\Pr[\mathsf{Verify}(pp, Q, \mathcal{A}(Q)) = 1 \wedge (Q, \cdot) \notin R] \leq \mathsf{negl}(\lambda)$$

Proof. Assume adversary $\mathcal{A}$ breaks soundness with non-negligible advantage ϵ. We construct algorithm $\mathcal{B}$ that solves DLP:

1. $\mathcal{B}$ receives DLP instance $(g, h = g^a)$
2. $\mathcal{B}$ simulates IFKP environment for $\mathcal{A}$, including all interactive reasoning cycles
3. When $\mathcal{A}$ produces valid proof π^*, $\mathcal{B}$ applies Multi-Model Extraction Lemma (Lemma 2.3)
4. If extraction produces commitment collision, $\mathcal{B}$ solves DLP
5. $\mathcal{B}$ succeeds with probability $\epsilon^2/\mathsf{poly}(\lambda)$

This contradicts DLP hardness. The interactive reasoning cycles ensure that any deviation from honest reasoning is either statistically detectable or requires breaking cryptographic commitments. □

2.5 Multi-Model Extraction Lemma

Lemma 2.3 (Multi-Model Extraction)**.** Let IFKP be the Full Knowledge Proof protocol. For any PPT adversary $\mathcal{A}$ that produces valid proofs with probability ϵ, there exists extractor $\mathcal{E}$ that either:

1. Extracts valid witness w for Q, or
2. Solves DLP in $\mathbb{G}$

with probability at least $\epsilon^2/\mathsf{poly}(\lambda)$.

Proof. The extractor $\mathcal{E}$ interacts with $\mathcal{A}$ using rewinding strategy across reasoning cycles:

1. For each reasoning cycle r and model i, $\mathcal{E}$ runs $\mathcal{A}$ multiple times to obtain multiple opening attempts

2. If any commitment opens to different messages $(y, r) \neq (y', r')$, then:

$$g^{H(y)}h^{H(r)} = g^{H(y')}h^{H(r')} \Rightarrow g^{H(y)-H(y')} = h^{H(r')-H(r)}$$

   Since $h = g^a$, we have $H(y) - H(y') = a \cdot (H(r') - H(r))$, allowing computation of a

3. If no collisions found, $\mathcal{E}$ uses valid openings to extract witness w across all reasoning cycles

By the forking lemma, success probability is $\epsilon^2/\mathsf{poly}(\lambda)$. The interactive reasoning cycles provide multiple independent opportunities for extraction. □

2.6 Statistical Security

Theorem 2.4 (Convergence Soundness)**.** Under model diversity and independence assumptions, the probability of false convergence across interactive reasoning cycles is bounded by:

$$\Pr[\mathsf{false\ convergence}] \leq e^{-O(k \cdot n \cdot R)}$$

where k is the outlier threshold, n is the ensemble size, and R is the number of reasoning cycles.

Proof. Follows from concentration inequalities and the statistical properties of median absolute deviation under independent sampling across multiple reasoning cycles. Each additional cycle exponentially reduces the probability of false consensus. □

3 Implementation

3.1 Reference Implementation

We provide a reference implementation for EIP-8004 integration featuring interactive reasoning cycles:

```
contract FKPValidator {
    struct FKPAttestation {
        bytes32 requestId;
        address agent;
        bytes32 claimHash;
        uint256 confidence; // basis points
        bytes32 reasoningTraceHash; // IPFS hash of all reasoning cycles
        bytes validatorSignature;
        uint256 timestamp;
        uint256 cyclesCompleted; // Number of reasoning cycles executed
    }

    mapping(bytes32 => FKPAttestation) public attestations;

    function validateRequest(
        bytes32 requestId,
        address agent,
        bytes calldata inputData
    ) external returns (bool approved, bytes memory proof) {
        // Off-chain FKP verification with interactive reasoning cycles
        FKPAttestation memory attestation = _runFKPVerification(
            requestId, agent, inputData
        );

        attestations[requestId] = attestation;
        bool isApproved = attestation.confidence >= MIN_CONFIDENCE;

        return (isApproved, abi.encode(attestation));
    }

    function challengeAttestation(
        bytes32 requestId,
        bytes calldata fraudProof
    ) external {
        _processChallenge(requestId, fraudProof);
    }

    function getReasoningCycles(bytes32 requestId) external view returns (uint256) {
        return attestations[requestId].cyclesCompleted;
    }
}
```

Listing 1: FKP Validator Contract with Reasoning Cycles

Table 1: Performance Characteristics of Interactive Reasoning Cycles

Operation	**Computation**	**Communication**	**Cost**
Initial Reasoning Cycle	$O(n)$ LLM queries	$O(n^2)$ messages	$0.10-$0.50
Additional Reasoning Cycles	$O(n \cdot R)$ queries	$O(n^2 \cdot R)$ messages	$1-$15
Verification	$O(1)$ exponentiations	$O(1)$ on-chain	$0.01
Challenge	$O(n \cdot R)$ recomputation	$O(1)$ on-chain	Slash amount

3.2 Performance Analysis

4 Applications

4.1 Autonomous AI Economies

FKP with interactive reasoning cycles enables trustless verification for:

- **AI Trading Agents**: Verify trading strategy reasoning across multiple market scenarios
- **DeFi Protocols**: Transparent risk assessment through multi-model reasoning cycles
- **Content Moderation**: Auditable content classification with reasoning transparency
- **Medical Diagnosis**: Verifiable diagnostic reasoning through specialist consensus

4.2 EIP-8004 Integration

FKP serves as a high-stakes trust model for EIP-8004's Validation Registry, providing stake-secured inference verification with full transparency of reasoning cycles.

5 Future Work

5.1 Limitations

- **Oracle Problem**: FKP verifies reasoning process integrity, not ground truth correspondence
- **Computational Cost**: Multiple interactive reasoning cycles require significant resources
- **Model Diversity**: Security depends on independent model training and reasoning approaches
- **Convergence Assumptions**: Some cognitive tasks may not admit clean probabilistic consensus

5.2 Future Directions

- Hybrid FKP-ZKP constructions for selective transparency in reasoning cycles
- Optimized convergence detection algorithms for complex reasoning tasks
- Cross-chain attestation verification for decentralized reasoning markets
- Formal verification of model independence across reasoning cycles

6 Conclusion

Full knowledge proofs represent a fundamental advancement in machine cognition verification, transforming opaque AI systems into transparent, auditable reasoning engines. By combining cryptographic commitments with multi-model interactive reasoning cycles and economic security, we create a trust layer suitable for the emerging autonomous economy.

The protocol's security rests on both computational hardness assumptions and statistical convergence properties across reasoning cycles, providing robust guarantees against coordinated manipulation. While not solving the fundamental oracle problem, FKP moves verification from ”trust the output” to ”verify the reasoning process”—a crucial step toward accountable autonomous systems.

The introduction of interactive reasoning cycles establishes a new paradigm for machine cognition verification, where truth emerges probabilistically through structured interaction rather than deterministically through isolated computation. This approach acknowledges the inherent uncertainty in complex reasoning while providing cryptographic guarantees about the consensus formation process.

As AI agents increasingly manage critical infrastructure and economic activity, full knowledge proofs provide the missing foundation for scalable trust, enabling the $30T autonomous economy while maintaining human oversight and accountability through transparent reasoning cycles.

Acknowledgments

We thank the AI research community and early supporters for their feedback and contributions. Our why.com platform serves as the live showcase for full knowledge proofs in production, demonstrating interactive reasoning cycles for verifiable machine cognition.

References

[GMR85] Goldwasser, S., Micali, S., & Rackoff, C. (1985). *The knowledge complexity of interactive proof-systems*. STOC '85.

[BFM88] Blum, M., Feldman, P., & Micali, S. (1988). *Non-interactive zero-knowledge and its applications*. STOC '88.

[BCC88] Ben-Or, M., et al. (1988). *Everything provable is provable in zero-knowledge*. CRYPTO '88.

[Ped91] Pedersen, T. P. (1991). *Non-interactive and information-theoretic secure verifiable secret sharing*. CRYPTO '91.

[Cra97] Cramer, R. (1997). *Modular design of secure yet practical cryptographic protocols*. PhD Thesis.

[Gro10] Groth, J. (2010). *Short pairing-based non-interactive zero-knowledge arguments*. ASIACRYPT 2010.

[BCI+13] Bitansky, N., et al. (2013). *Recursive composition and bootstrapping for SNARKs and proof-carrying data*. STOC '13.

[BBHR18] Ben-Sasson, E., et al. (2018). *Scalable, transparent, and post-quantum secure computational integrity*. IACR Cryptology ePrint Archive.

[BGKW88] Ben-Or, M., Goldwasser, S., Kilian, J., & Wigderson, A. (1988). *Multi-prover interactive proofs: How to remove intractability assumptions*. STOC '88.

[BFL91] Babai, L., Fortnow, L., & Lund, C. (1991). *Non-deterministic exponential time has two-prover interactive protocols*. Computational Complexity.

[Val08] Valiant, P. (2008). *Incrementally verifiable computation or proofs of knowledge imply time/space efficiency*. TCC 2008.

[BSCG+13] Ben-Sasson, E., et al. (2013). *SNARKs for C: Verifying program executions succinctly and in zero knowledge*. CRYPTO 2013.

[Soa16] Soares, N. (2016). *The value learning problem*. Alignment Research Center.

[Rus19] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

[LBD+89] LeCun, Y., et al. (1989). *Backpropagation applied to handwritten zip code recognition*. Neural Computation.

[CPA+21] Perez, E., et al. (2021). *Red teaming language models with language models*. arXiv:2102.06720.

[Die00] Dietterich, T. G. (2000). *Ensemble methods in machine learning*. MCS 2000.

[JSS90] Jacobs, R. A., et al. (1990). *Adaptive mixtures of local experts.* Neural Computation.

[GVR17] Gunning, D., Vorm, E., & Riedl, M. O. (2017). *Explainable AI: The basics.* DARPA.

[RSG16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier.* KDD 2016.

[LL17] Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions.* NeurIPS 2017.

[KLV+22] Kouvaros, P., Lomuscio, A., & Vantaggiato, M. (2022). *Formal verification of neural networks: A survey.* ACM Computing Surveys.

[EZ19] Ellis, S., & Zargham, M. (2019). *Chainlink: A decentralized oracle network.* Chainlink Whitepaper.

[Fet19] Fetch.ai (2019). *Fetch.ai: Building the decentralized digital world.* Fetch.ai Whitepaper.

[OC21] Ocean Protocol (2021). *Ocean Protocol: Tools for the Web3 data economy.* Ocean Protocol Whitepaper.

[Bit21] Bittensor (2021). *Bittensor: A peer-to-peer intelligence market.* Bittensor Whitepaper.

[ZGK+23] Zhang, Y., et al. (2023). *zkML: Verifiable machine learning with zero-knowledge proofs.* IEEE S&P 2023.

[GDL+21] Gilad-Bachrach, R., et al. (2021). *CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy.* ICML 2016.

[LSP82] Lamport, L., Shostak, R., & Pease, M. (2022). *The Byzantine Generals Problem.* ACM Transactions on Programming Languages and Systems.

[CL99] Castro, M., & Liskov, B. (1999). *Practical Byzantine Fault Tolerance.* OSDI '99.

[Nak08] Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system.*

[KZG16] Kiayias, A., et al. (2016). *Ouroboros: A provably secure proof-of-stake blockchain protocol.* CRYPTO 2017.

[Get63] Gettier, E. L. (1963). *Is justified true belief knowledge?* Analysis.

[Ben99] Benkler, Y. (1999). *Free as the air to common use: First Amendment constraints on enclosure of the public domain.* NYU Law Review.

[Pfi99] Pfitzmann, A. (1999). *Multilateral security: Enabling technologies and their evaluation.* IFIP SEC.

[CF13] Catalano, D., & Fiore, D. (2013). *Vector commitments and their applications.* PKC 2013.

[KZG10] Kate, A., Zaverucha, G., & Goldberg, I. (2010). *Constant-size commitments to polynomials and their applications.* ASIACRYPT 2010.

[GGP10] Gennaro, R., Gentry, C., & Parno, B. (2010). *Non-interactive verifiable computing: Outsourcing computation to untrusted workers.* CRYPTO 2010.

[GGI+22] Gentry, C., et al. (2022). *Pinocchio: Nearly practical verifiable computation.* IEEE S&P 2013.

[BM88] Bellare, M., & Micali, S. (1988). *How to sign given any trapdoor function.* STOC '88.

[BFLS91] Babai, L., Fortnow, L., Levin, L. A., & Szegedy, M. (1991). *Checking computations in polylogarithmic time.* STOC '91.

[BK23] Borge, M., & Khabbazian, M. (2023). *Proof of necessary work: Succinct state verification with fairness guarantees.* FC 2023.

[Gartner2030] Gartner Research (2023). *Autonomous AI Economy Projections.* Gartner Research.