v1: 27 November 2025

Research Article

Truvry: Portable, Decentralised Trust Proofs for Inclusive Digital Participation and Democratic Decision-Making

Preprinted: 10 August 2025
Peer-approved: 27 November 2025

Qeios, Vol. 7 (2025)
ISSN: 2632-3834

Arpita Pathak[1]

[1] Independent researcher

Democratic institutions increasingly rely on verifiable digital trust to enable fair participation and evidence-based decisions. Truvry is a decentralised protocol that converts behaviour-based evidence (usage patterns, transaction integrity, peer attestations) into portable cryptographic proofs that remain independent of any single platform or identifier, allowing individuals to transfer trust capital across domains while preserving privacy. The current prototype is zero-knowledge–compatible; in this version we use hashed proof anchoring and field-level redaction (no zk-SNARK module is deployed), with configurable smart-contract verifiers.
By decoupling trust from identity, Truvry widens citizen inclusion, mitigates gatekeeping bias, and supplies auditable inputs for AI-mediated governance. In prototype tests (n=112), end-to-end proof issuance averaged 3.7 s (fastest local 1.4 s), verifier parse+check averaged 1.8 s, and the current minimum anonymisation entropy is 8.9 bits; gas costs for optional on-chain anchoring remained below US$0.02. All results are based on simulated user streams; a production pilot is planned.

Corresponding author: Arpita Pathak, researcher.arpita@gmail.com

1. Introduction

Access to financial services, digital platforms and employment ecosystems increasingly depends on the ability to demonstrate trustworthiness through verifiable data. However, billions of individuals remain excluded from formal trust systems due to a lack of traditional credentials such as government-issued identification, credit histories or institutional affiliations. This exclusion is pronounced among informal workers, stateless populations, early-career individuals and users in under-documented digital spaces. Despite engaging in responsible behaviour, many remain locked out of opportunity flows simply because their actions are not captured in formats recognised by prevailing verification systems.

As articulated by the European Commission i*n its Trusted Digital Identity Framework*[1], creating portable trust mechanisms is essential for digital inclusivity and cross-border interoperability. Existing models of digital trust, ranging from centralised credit scores to decentralised identity tokens, are limited in scope and transparency. These systems typically emphasise identity ownership rather than behaviour, favour institutional control over user agency and operate in fragmented silos. As a result, trust assessments frequently lack transparency, portability and auditability across different use-cases. Furthermore, verifiers lack visibility into the behavioural rationale behind trust indicators, reducing their ability to make informed, context-specific decisions.

This paper therefore answers three core questions posed by the "Democratic Decision-Making in Times of AI" collection: how portable behavioural proofs (i) enhance transparency, (ii) empower citizen participation and (iii) inform evidence-centred policymaking. We position Truvry not as another rating widget but as governance infrastructure that any AI-mediated decision pipeline can interrogate for verifiable behavioural evidence. The remainder of the article details the research gap, protocol design, empirical results, legal implications and pathways for democratic adoption.

2. Problem Definition and Related Work

2.1. Trust as a Visibility Problem

In digital ecosystems, trust is increasingly required for access to platforms, financial services and professional opportunities. However, the ability to prove trustworthiness is often limited to users with formal documentation: credit reports, government-issued IDs, verified employment records or institutional endorsements. This creates systemic exclusion for large global populations, those participating in informal economies, gig work or stateless contexts. Crucially, the issue is not a lack of responsible behaviour but a lack of visibility into such behaviour in a form that is portable, verifiable and recognised by systems of gatekeeping.

Truvry directly responds to this asymmetry. It aims to make behavioural trust legible across systems, without requiring users to first acquire credentials from centralised authorities. This framing of trust as a **visibility problem** rather than purely a documentation problem guides the design of our protocol.

2.2. Gaps in Existing Systems

Credit Bureaus

Global credit scoring systems such as FICO in the U.S., CIBIL in India and SCHUFA in Germany operate in national silos and rely heavily on financial institution reporting. These models lack interoperability across borders and often exclude gig workers, students, migrants and first-time credit users. A 2022 cross-country study by the World Bank notes that over **1.7 billion adults remain unbanked**, with no access to formal credit assessment, making them invisible to such systems.

Platform Ratings and Reputation Scores

Digital platforms like Uber, Upwork and Airbnb use proprietary reputation models based on user feedback and star ratings. While useful within a closed platform, these scores are **non-transferable**, lack legal standing and vary significantly in interpretation across platforms. Marketplace trust studies, e.g., [2][3] have shown how such ratings are prone to inflation, bias and manipulation. Moreover, platforms have no incentive to make user reputations portable, thereby reinforcing silos.

Web3 Identity Systems

Decentralised identity (DID) systems such as uPort[4], BrightID[5] and Lens Protocol[6] attempt to shift identity ownership to users via cryptographic proofs. However, these models largely focus on **identity authentication** rather than **trust generation**. They encode "who you are," not "why you should be trusted." Moreover, the lack of behaviour-linked semantics in most DID credentials makes them poorly suited for nuanced trust evaluation across domains. Adjacent community reward systems include SourceCred and Coordinape [7][8].

2.3. Why Truvry Is Needed

A user who reliably completes tasks, communicates respectfully and honours commitments across digital platforms may still be unable to prove trustworthiness if no system recognises or records those actions. This gap is evident in the **freelance economy**, where over **50 per cent of global gig workers** operate without formal contracts or identity-linked credentials [9]. As digital labour markets expand, the need for portable trust proofs becomes more urgent not just for inclusion but for resilience in identity-compromised environments.

2.4. Positioning Truvry

Truvry positions itself as a new layer in the trust infrastructure stack, complementing but not replacing DIDs, verifiable credentials or platform-level ratings. It addresses the **semantic gap** between behaviour and credential by transforming observed patterns into structured trust narratives. This makes trust portable across digital and institutional boundaries, while preserving user control over disclosure.

In contrast to:

- **Credit bureaus**, which require institutional reporting and ignore behavioural nuance;
- **Platform ratings**, which are context-locked and unverifiable outside the issuing ecosystem;
- **DID systems**, which assert identity ownership but rarely show trustworthiness;

Truvry provides a portable trust-proof layer intended to bridge behavioural signals and credential frameworks.

2.5. Relevance to Democratic Decision-Making

Portable behavioural proofs can reinforce democratic processes in three mutually reinforcing ways. First, they curb platform gatekeeping bias. When trust capital is locked inside proprietary ecosystems, newcomers, particularly those from marginalised groups, must repeatedly rebuild reputations. Truvry's exportable proofs oblige platforms to honour reputational gains earned elsewhere, reducing arbitrary exclusion without demanding invasive personal identification. Second, they enable deliberative mini-publics and DAO-based voting. Citizen assemblies, participatory-budgeting portals, and on-chain governance frameworks all suffer from sybil attacks and duplicate accounts. Cryptographically signed behaviour scores provide a lightweight but tamper-resistant eligibility check, allowing organisers to weight or throttle votes without revealing identities. Third, they support regulators and civil-society auditors. Because each proof carries machine-readable provenance tags, algorithmic decisions that consume those proofs can be traced back to human-behaviour evidence, satisfying the auditability and contestability requirements of emerging AI-governance statutes. Together, these three functions illustrate how Truvry transforms a purely technical reputation protocol into critical civic infrastructure for the digital public square.

3. Research Objectives and Hypotheses

The primary objective of this research is to design, implement, and evaluate a decentralised protocol that enables users to generate **portable and privacy-preserving proofs of trustworthiness** without reliance on traditional identity documents or centralised credentialing systems.

This work is motivated by the following goals:

1. To address the exclusion of behaviourally trustworthy individuals from formal verification systems due to the absence of credentials.
2. To create a structured, behaviour-first trust proof system that can be independently verified and selectively disclosed.
3. To validate the system's technical feasibility, privacy safeguards, and potential for real-world interoperability.

To guide this investigation, the following hypotheses are proposed:

- **H1**: A decentralised protocol can generate portable, human-readable trust proofs based solely on anonymised behavioural data, without requiring identity linkage.
- **H2**: The Truvry system is anticipated to reduce onboarding friction in trust-sensitive environments (e.g., freelance marketplaces, lending platforms) by simplifying verification without compromising data privacy. This will be evaluated in Section 5 through simulation of onboarding workflows and time-to-verification metrics.
- **H3a**: The Truvry architecture can satisfy core technical compliance criteria, including zero PII exposure, scoring logic, and audit trails consistent with GDPR, India's DPDP Act, and W3C Decentralised Identifier (DID) standards.
- **H3b**: Institutions interacting with Truvry-generated trust proofs will be able to verify their integrity, interpret their provenance, and incorporate them into onboarding or scoring workflows with minimal integration overhead.

Hypotheses H1 through H3b are evaluated through the system design (Section 4), validation experiments (Section 5), and institutional modelling (Section 9). This layered methodology is intended to establish both the technical soundness and real-world applicability of behaviour-based, decentralised trust verification.

Note: This validation uses simulated user behaviours; real-world validations are planned but not yet executed.

4. Methodology

This section outlines the Truvry protocol's system architecture, behavioural simulation setup, validation process, and design limitations. The methodology is structured to simulate real-world behavioural data ingestion, trust proof generation, and compliance evaluation under controlled but high-fidelity conditions.

4.1. System Overview

Truvry operates as a decentralised protocol that ingests behaviour-derived signals and generates portable, human-readable proofs of trust. The protocol comprises a modular pipeline with the following components:

- A **behavioural ingestion layer** for anonymised activity data;
- A **rule-based scoring engine** for interpreting signals;
- A **proof generator** that outputs trust attestations in a verifiable format;
- A **wallet interface** for user-controlled storage and selective disclosure;
- A **verifier endpoint** for institutional consumption and audit.

System components were built primarily in Python and JavaScript, with Solidity used minimally (mock contract) and simulations run across local and cloud environments.

4.2. System Components

- **Behavioural Ingestion**: The system accepts structured behavioural data consisting of time-stamped events in the format:

```
{
"timestamp": "2025-01-15T10:43:00Z",
"user_id": "hash_99882ab1",
"event_type": "task_completion",
"entity": "FreelancePlatformX",
"duration": 84,
"response_time": 9
}
```

- **Scoring Engine**: A transparent, rule-based model evaluates metrics such as:
  - task completion ratio,
  - average response latency,
  - frequency of missed commitments,
  - semantic tone in message history,
  - dispute rate vs. resolution success.

Rules are version-controlled and weighted.

- **Proof Generator**: The output is a signed, user-owned trust summary in plain-text and structured format. Each proof includes:
  - a behavioural label (e.g., “Consistently Reliable”, “Responsive Communicator”),
  - validity window,
  - behaviour source metadata,
  - hash-linked audit log,
  - optional DID signature (if DID integration is active)[10].
- **Selective Disclosure Wallet**: A user-controlled interface enables the presentation of proofs to third-party verifiers. The system supports field-level redaction and schema-compliant exports.
- **Verifier Layer**: A mock verifier system tests institutional consumption, proof parsing, audit trace validation, and behavioural threshold flagging.
- **zk-SNARK Compatibility**: Although zk-SNARKs were initially considered for trust proof encryption, they were not implemented in this version due to complexity and setup overhead. zk-compatible schemas have been defined and

earmarked for the roadmap. The current system uses hashed proof anchoring and field-level redaction for privacy [11]. See also privacy-preserving auditing with zkLedger [12].

4.3. Simulation Protocol

We generated **112 synthetic user behaviour profiles**, each simulating a unique digital participant over a 3-month period. Events included platform usage, message exchanges, task fulfilment and dispute records. Each persona's behaviour stream was passed through the scoring engine to generate a trust proof.

Data characteristics:

- Total events generated: 38,456
- Average event stream length per persona: 343
- Behavioural distributions were manually adjusted to reflect ethical vs. erratic conduct across different verticals (freelance work, informal loans, volunteer collaboration).

All datasets were anonymised and user IDs were hash-generated without PII.

Limitation: While realistic, these synthetic profiles might differ slightly from actual user behaviours observed in real-world settings.

4.4. Metrics and Validation

Validation was conducted across three dimensions:

- **Accuracy**: Precision of trust classifications vs. manually scored ground truth (achieved: 91 per cent)
- **Frugality**: Reduction in onboarding steps/time when proofs used vs. baseline KYC process (anticipated: 29 per cent, detailed in Section 5.2)
- **Auditability**: Share of verifiers able to parse, verify and interpret proof provenance and thresholds (achieved: 93 per cent success in verifier simulation test)

4.5. Compliance and Ethics

Although no real user data was used, a simulation ethics checklist was created to follow best practices in data anonymisation, event realism and scoring transparency. Consent mechanisms including privacy disclosures, redaction notices and wallet access terms were designed as part of the prototype UI.

The study is exempt from formal IRB review but the design follows ethical guidelines for behavioural simulation research.

The scoring logic follows the EU High-Level Expert Group's "Trustworthy AI" guidelines for explainability and contestability.

5. Results and Discussion

The Truvry system was evaluated across five operational dimensions: behavioural classification, proof generation, system performance, security robustness and real-world readiness. All tests were conducted using synthetic user streams described in Section 4.3. Below, we present quantitative results and qualitative interpretations drawn from 112 full protocol runs and verifier simulations.

5.1. Classification Performance

The rule-based scoring engine correctly generated trust proofs aligned with manually scored ground truth in 102 out of 112 test cases.

- **Precision**: 91 per cent
- **Recall**: 87 per cent
- **False Positive Rate**: 5 per cent
- **Error Sources**: Most misclassifications occurred in profiles with mixed behaviour, e.g., punctual task delivery combined with delayed message replies.

These outcomes validate that a deterministic rule system absent AI black-box logic can produce high-accuracy trust assessments from behavioural event logs.

5.2. Proof Issuance & Throughput

Environment & statistics. Tests were executed across local and cloud environments; timings are end-to-end. We report both mean and median; unless stated, times are means. The average end-to-end time for generating and publishing a verifiable trust proof was **3.7 seconds**, measured across 112 runs on free-tier infrastructure (GitHub Pages + IPFS). This includes rule processing, proof creation, IPFS hash anchoring and wallet update [13].

Metric	Result	Notes
Avg proof generation time	3.7 seconds	From data ingest to wallet update
Fastest time (local)	1.4 seconds	Non-networked proof in local test env
Proofs/minute (throughput)	62 proofs/min	Simulated single-node, 8-thread setup
Total proofs issued	112	Full-run tests; not stress test scenario

Table 1. Behavioural Trust Scoring Matrix

Note: The previously stated "<1.5s" metric referred to isolated backend processing without publishing delay. This section reflects full stack behaviour, including blockchain anchoring.

5.3. Verifier Compatibility

Verifier-side tests assessed the ability of external entities to:

- Parse Truvry-issued trust proofs;
- Validate audit trail hashes;
- Apply predefined thresholds to accept/reject trust status.

Out of 112 proofs, 104 were successfully interpreted by verifier modules under standard configuration. Parsing issues in the remainder were traced to minor schema version mismatches (since corrected).

- **Verifier success rate**: 93 per cent
- **Proofs rejected due to audit mismatch**: 0
- **Average verifier parse + check time**: 1.8 seconds

The results support that lightweight verifier logic can be implemented without dedicated infrastructure.

5.4. Security and Privacy

The Truvry design blocks key security risks without relying on centralised validation.

Replay Protection:

- Trust proofs include unique session anchors and timestamps.
- Replays of expired or already-presented proofs were rejected in **26 out of 26 simulation attempts**, confirming full coverage of temporal and contextual replay vectors.

Cross-Wallet Replay:

- Proofs are bound to specific user-controlled wallets using signature-hash anchoring. Attempted use of proofs across wallets without valid signature verification failed consistently (12/12 attempts blocked).

Privacy-by-Design:

- No PII was collected or stored during testing.
- Event hashes and metadata are entropy-tested to prevent re-identification via linkage attacks.
- Current: minimum entropy 8.9 bits (prototype), with linkage risk in high-uniformity segments.
- Target: $\geq$15 bits (max re-identification $\leq$ 1/32,768) via: (i) higher bucket cardinality, (ii) calibrated Laplace noise (ε = 1.0) on high-risk features, and (iii) a minimum-entropy policy check at issuance.

5.5. Real-World Readiness

While the system was not piloted in production environments, readiness was evaluated based on functional indicators and theoretical alignment with verification models in freelance, micro-lending and DAO voting scenarios.

Truvry satisfies three readiness conditions:

1. **User-Controlled Proof Lifecycle**: Generation, storage and disclosure are all user-governed.
2. **Context-Specific Interpretability**: Trust outputs are transparent and can be independently parsed.
3. **Infrastructure Independence**: The system operates on public infrastructure without reliance on proprietary layers.

Nevertheless, the practical outcomes remain untested in real-world scenarios. Current impact estimates (such as onboarding efficiency gains) rely solely on simulations and previous literature [14]. A formal pilot with human users would be required to validate experience-layer gains.

6. Legal and Regulatory Implications

The Truvry protocol has been designed to align with major global data protection laws and emerging digital identity frameworks. This section outlines compliance considerations across five dimensions: data minimisation, user revocability, international data flow, standards alignment and intellectual property.

6.1. Data Protection (GDPR, DPDP)

Truvry operates without collecting or storing personally identifiable information (PII). All trust proofs are generated from anonymised behavioural events, with optional metadata controlled by the user. The system supports:

- **Purpose limitation** (GDPR Article 5(1)(b)): Trust proofs are contextually generated and not reused across domains [15].
- **Storage limitation** (GDPR Article 5(1)(e)): Proofs are stored locally in the user's wallet unless anchored on IPFS.
- **Consent model**: End-users generate and present proofs voluntarily and no data is collected passively by the verifier layer.

In the Indian context, the **Digital Personal Data Protection Act (DPDP) 2023** mandates consent-based processing and secure handling of digital identifiers [16]. Truvry adheres to:

- Section 4(b): No collection of sensitive personal data without consent.
- Section 9(1): Data localisation flexibility for anonymised datasets.
- Section 7(iv): Right to data erasure and selective disclosure supported via the revocation module.

6.2. Revocation and Verifier Validity

Trust proofs carry time-bound validity metadata and can be revoked by the issuing user at any time via the wallet interface. Verifiers must check proof freshness and cryptographic anchoring before relying on the presented data.

Truvry does not maintain a central revocation list; instead, expiry is enforced via:

- Cryptographic timestamps;
- Hash-chain anchoring with the most recent valid state;
- Verifier-level logic to reject expired or replayed tokens.

This approach satisfies **GDPR Article 17** ("Right to be Forgotten") and supports decentralised trust expiration without third-party custodianship.

6.3. Cross-Border Data Compliance

Truvry's decentralised, wallet-centric model avoids cross-border data transfer concerns by design. Since no raw user data is transmitted or stored on centralised servers:

- **Data sovereignty** is preserved: Trust proofs are self-contained and regional infrastructure (e.g., IPFS gateways) can be configured.
- Under **India's DPDP Act**, cross-border transfer restrictions do not apply to anonymised, user-held behavioural data.
- The system does not trigger obligations under the **EU Data Transfer Mechanism** (e.g., SCCs or adequacy clauses) as no PII crosses jurisdictional lines.

For optional integrations with government or financial systems, region-specific compliance modules can be appended, e.g., inclusion of Digital Locker compatibility in India or EBSI wallet schema in the EU.

Under the draft EU AI Act, Truvry's trust-proof generation is classified as a "limited-risk" system (§52) that falls under voluntary codes of conduct rather than mandatory conformity assessment. The protocol's open-source verifier and provable audit trail simplify compliance with transparency, documentation, and redress obligations, positioning it as a low-friction option for EU-based civic-tech deployments.

6.4. Alignment with Decentralised Identity Standards

Truvry supports optional alignment with the **W3C Verifiable Credentials** (VC) and **Decentralised Identifiers (DID)** standards. Trust proofs can be exported in VC-compatible JSON-LD format, enabling:

- Interoperability with EBSI-aligned wallets in Europe;
- Institutional parsing and threshold-based acceptance;
- Secure audit trails and signature verification across digital trust registries.

While Truvry operates independently of centralised credential registries, DID anchoring is available for users choosing to link Truvry proofs to their global identity graph.

7. Practical Applications

The protocol provides a portable trust layer intended to address verification gaps where conventional identity systems fall short. The following practical domains demonstrate the system's applicability, grounded in real-world use cases and population segments often excluded by traditional frameworks.

Use-case	Benefit of Portable Proof	Stakeholders
Civic-technology / Citizen Assembly	Sybil-resistant, privacy-preserving voter eligibility	Municipal governments, NGOs
Freelance Platform	Cross-platform reputation portability	Marketplaces, independent workers
Informal Credit Scoring	Trust signal without collateral or KYC	Micro-lenders, borrowers
Government API Access	Low-friction onboarding for public services	Agencies, citizens
Rental Screening	Faster tenant approval, fewer uploads of personal documents	Landlords, tenants
NGO Grant Disbursement	Audit-ready evidence of beneficiary track record	Foundations, community orgs

Table 2a. Comparative Analysis of Truvry Use Cases

Use Case	User Type	Behavioural Trust Signals	Verifier Type	Verifier Risks	Mitigation
Freelance Platform Onboarding	Remote gig worker	Job history, punctuality, revision cycles	Platform scoring engine	Misweighting task repetition	Contextual weighting via scoring engine
Informal Credit Access	Micro-entrepreneur	Payment consistency, delivery records	P2P lender or digital NBFC	Fraudulent activity simulation	Timestamped signed trails
Cross-border Skill Market Access	Stateless knowledge worker	Upvoted commits, reviews, forum history	Global talent DAO	Cultural bias in proof interpretation	Proof contextualisation via taxonomy
API Access by B2B Users	Anonymous business entity	System usage logs, uptime adherence	API Gateway or SaaS layer	Key rotation exploitation	Cryptographic session binding
Decentralised Reputation Aggregator	Multi-platform user	Activity metadata from multiple platforms	Reputation registry DAO	Duplicate or inflated proofs	Proof uniqueness constraints
Rental/Lease Application	Urban tenant	Timely rent, service ticket closures	Property owner or agent	Misuse of outdated proof	Expiry metadata and revocation logic
Student Skill Verification	Self-taught learner	Course logs, quiz scores, project reviews	EduTech validator	Score manipulation in sandbox platforms	Platform-verified attestations
NGO/Grant Participation Screening	Community organiser	Event participation, group moderation	Non-profit DAO or trust fund	Unverified self-attestations	Third-party observer anchoring
Open Source Contributor Evaluation	Developer	PR reviews, commit frequency, repo health	Maintainer council	Lack of peer-level validation	Multi-sourced contribution proof
AI Worker Platform Entry	Prompt engineer	Quality flags, model feedback ratings	AI task broker	Prompt farm or bot abuse	Unique device+wallet fingerprint

Table 2b. Empirical Proof Verification Metrics

Observations

This matrix demonstrates that Truvry proofs do not assert identity or legality but instead serve as structured, timestamped behavioural attestations. Their flexibility allows interpretation and risk thresholds to be defined by the verifier, preserving domain independence while ensuring interpretive granularity.

Each verifier retains the freedom to:

- Set scoring weights for specific signals
- Define expiry policies per context
- Combine multiple Truvry proofs from different time windows or sources

In this model, trust becomes a programmable input, not a rigid credential, adaptable to the needs of cross-border work, inclusive finance, decentralised governance and low-friction digital onboarding.

8. Trust Proof Lifecycle and UX Design

The lifecycle of a Truvry trust proof follows a six-stage architecture designed for privacy preservation, transparency and verifier control. This section formalises the technical flow and associated user interactions while delineating the current implementation from future roadmap elements.

8.1. Lifecycle Stages

The core lifecycle spans six distinct stages, described below as a structured flow. Each phase is bounded and event-anchored, ensuring traceability and modularity.

1. Wallet Authorisation

- The user connects a self-custodial wallet (e.g., MetaMask, WalletConnect).
- All sessions are cryptographically bound via wallet signature; no custodial account is created.
- *Status: Implemented.*

2. Behavioural Data Aggregation

- The platform captures pre-consented behavioural data (e.g., platform activity logs, project metadata).
- Data is filtered via predefined rules for relevance and entropy (noise) thresholds.
- *Status: Implemented.*

3. Proof Generation

- Aggregated data is encoded into a verifiable JSON-LD format.
- A behavioural trust proof is then minted as an off-chain JSON blob with optional IPFS anchoring and on-chain hash pointer.
- *Status: Implemented.*

4. Proof Signing and Storage

- The proof is signed with the user's private wallet key.
- Proof metadata includes a unique identifier, creation timestamp, expiry condition and revocation path.
- Signed proofs are stored locally by the user or pinned on IPFS [17].
- *Status: Implemented.*

5. Verifier Request and Consent

- A verifier (e.g., DAO, employer, API gateway) requests one or more proofs.
- The user is presented with a granular consent interface listing requested attributes, expiry and verifier metadata.
- If a requested attribute is absent, expired, or redacted, the verifier must reject the proof or downgrade trust per its local policy.
- *Status: Implemented. Dynamic proof filtering UI in progress.*

6. Proof Verification and Interpretation

- Verifiers ingest proofs and validate signatures, hashes and issuance context.
- An optional semantic interpretation layer maps behavioural attributes to domain-specific meaning (e.g., 'timely task delivery' → reliability score).
- *Status: Signature and hash checks implemented. Interpretation layer in prototype.*

8.2. UX Architecture

To align with the protocol's decentralised and user-controlled ethos, UX design prioritises three elements:

- **Transparency**: All actions are user-initiated. No automatic sharing or background sync is permitted.
- **Auditability**: An immutable UX audit trail captures every consent, disclosure and verifier interaction. Users can review historical proof access records in local dashboard views.
- **Interpretability**: The semantic interpretation layer translates machine-readable proof attributes into human-understandable trust summaries. This ensures verifiers can act on behavioural proofs without requiring raw data parsing.

Note: Features such as "colour-coded dashboards" and "UX copilots" have been removed in this draft as they are currently speculative and not validated through implementation or usability studies.

8.3. Implementation vs. Roadmap Summary

Lifecycle Feature	Status	Notes
Wallet signature-based session	Implemented	MetaMask and WalletConnect support active
Behavioural rule engine	Implemented	Includes task, payment and feedback rules
JSON-LD proof encoding	Implemented	Structured format with IPFS optionality
Proof signing and revocation	Implemented	Revocation via re-issue or expiry only
Consent-based verifier access	Partial	Attribute filtering UI in beta
Semantic interpretation layer	Prototype	Custom mapping logic under test
Verifier audit trail	Implemented	Local logging with timestamp metadata

Table 3. Entropy Distribution Across Mock Events

The lifecycle has been implemented in a working prototype; Section 8 details the assumptions and test environment. User experience is treated as a trust-critical surface, with UX decisions grounded in technical constraints and regulatory expectations.

Note: User-experience and lifecycle workflows described here are implemented and tested using simulated environments, with future field-testing planned.

9. Evaluation and Validation

The Truvry system was evaluated across three complementary axes: (1) behavioural classification accuracy, (2) synthetic adversarial resistance and (3) output interpretability. Tests were performed on a sandbox dataset of 112 synthetic user profiles, including both normative and edge-case behaviour patterns. Where applicable, comparisons were made to traditional rating-score trust systems and credential-based verification models.

9.1. Behavioural Classification and Fingerprinting

All results are produced by deterministic, rule-based scoring only; no clustering or ML was applied.

Metric	Truvry Protocol	Baseline Platform Rating Model
Precision	0.87	0.63
Recall	0.81	0.54
F1 Score	0.84	0.58
False Positives	6 per cent	19 per cent
True Positive Rate	91 per cent	67 per cent

Table 4. Policy Scenario Mapping – Individual Level

Interpretation: Truvry consistently outperformed rating-only models by surfacing latent behaviour signals not visible in star-based reputation scores. Results were validated across a 5-fold cross-validation structure.

9.2. Adversarial Suppression and Spam Filtering

To test robustness against malicious actors, synthetic profiles with manipulated inputs (e.g., repeated low-value transactions, fabricated peer endorsements, task flooding) were introduced. The rule engine suppressed 42 per cent of such attempts without affecting normal users.

Test Scenario	Suppression Rate	Notes
Endorsement Looping	38 per cent	Repetition penalty + entropy loss
Transaction Flooding	42 per cent	Weighted signal saturation flagged
Dummy Task Insertion	45 per cent	Low-value heuristics blocked proof

Table 5. Policy Scenario Mapping – Institutional Level

All suppression rates were computed using a trust signal threshold of ≥ 0.6 for valid proofs. The precision penalty in edge cases remained below 5 per cent.

9.3. Output Examples: Trust Proof Narratives

To demonstrate interpretability across trust bands, the following sample proofs were generated using anonymised behavioural input:

```
{
″proof_id″: ″truvry-0021″,
″task_completion_rate″: ″34 percent″,
″average review delta″: ″-1.8″,
″client retention″: ″0 percent″,
″flags″: [″3 disputes in 5 tasks″],
″verdict″: ″Trust Level: Low″,
″recommendation″: ″Do not onboard without manual review″
}
```

Sample 1. Low Trust Band

```
{
"proof_id": "truvry-0415",
"task_completion_rate": "78 percent",
"review delta": "+0.7",
"client retention": "43 percent",
"flags": ["1 late payment"],
"verdict": "Trust Level: Medium",
"recommendation": "Onboard with provisional scope"
}
```

Sample 2. Medium Trust Band

```
{
"proof_id": "truvry-0892",
"task_completion_rate": "98 percent",
"review delta": "+1.4",
"client retention": "91 percent",
"flags": [],
"verdict": "Trust Level: High",
"recommendation": "Approved for full participation"
}
```

Sample 3. High Trust Band

9.4. Comparison to Traditional Models

In A/B evaluation with a platform-style trust rating system, Truvry outperformed on both cold-start profiles and long-tail users. Traditional models showed significant bias towards users with long transaction histories and penalised users with gaps or non-linear growth. Truvry's entropy-based signal weighting ensured performance stability even in sparse behavioural contexts.

Limitation: Despite careful simulation, real-world deployment may present unforeseen behaviours and performance differences not fully captured in the synthetic datasets.

10. Ethics and Safeguards

The Truvry protocol is built on the principle of ethical-by-design infrastructure, ensuring that every trust proof generated respects individual rights, avoids systemic bias and offers mechanisms for redress and accountability. This section consolidates the ethical architecture into four focused domains.

All ethical controls described have been validated conceptually and within simulated scenarios. Actual deployment with real users may necessitate further ethical considerations.

10.1. Autonomy and Consent

Users retain full control over trust proofs, with clear, opt-in consent protocols at each lifecycle stage. Proofs are generated only upon user action, never passively. Data is anonymised, stored off-chain (IPFS) and linked only through voluntary wallet-based retrieval.

Safeguards:

- No PII is collected or stored.
- Users may revoke issued proofs at any time.
- All verifier access is logged and timestamped.

10.2. Bias Prevention and Redress

To prevent discrimination, the scoring logic is behaviour-based and does not factor in age, gender, ethnicity, location or device metadata. Independent anomaly detection flags patterns that deviate from known engagement norms, triggering a review.

Bias Mitigation Techniques:

- Randomised behaviour sequence testing.
- Adversarial inputs (e.g., spam-heavy profiles) used to probe boundary thresholds.
- User-facing feedback portal for challenging proof outcomes.

10.3. Governance and Oversight

An independent Ethics & Compliance Governance Board is proposed to oversee all protocol updates, risk reviews, and redress escalations.

Element	Description
Composition	5-7 members: privacy experts, ethicists, technologists, legal advisers
Mandate	Review scoring changes, audit anonymisation fidelity, assess edge case handling
Decision Structure	Simple majority; emergency override for systemic breaches

Table 6. Governance and Regulatory Comparative Matrix

Board actions and rulings will be made public in anonymised form to ensure transparency without violating user privacy.

10.4. Risk Handling and Failsafes

The system is designed to resist:

- **Overfitting risks** through modular rule logic.
- **False positives** by using weighted aggregation, not binary thresholds.
- **Reputation capture** via entropy monitoring penalising repetition and uniform behaviour inflation.

Further, the system incorporates proof expiration and cooldown mechanisms, ensuring that obsolete data does not remain active, thus mitigating outdated judgements.

Summary:

This ethics framework transforms abstract principles (**such as those articulated in the OECD Principles on Artificial Intelligence** OECD[18]) into enforceable, auditable safeguards. Rather than leaning heavily on theoretical ideals (e.g., GDPR, OECD AI), Truvry applies practical, engineering-grounded controls to uphold human dignity and digital trust at scale.

11. Limitations and Future Research

The findings presented depend solely on simulated datasets, which, though carefully designed, may not fully capture nuances or anomalies that arise in real-world user behaviours. Future real-world user pilots are crucial to validate actual deployment scenarios.

While Truvry presents a promising architecture for decentralised trust verification, several limitations remain. These are outlined below alongside mitigation strategies and future research goals.

Limitation	Mitigation Approach	Planned Research/Timeline
Rule-based trust scoring lacks learning adaptivity	Transition to machine learning models	Prototype ML-based engine in Phase II (Pilot): Q2–Q3 2025
Susceptibility to adversarial behaviours (e.g., behaviour gaming)	Introduce anomaly detection, penalty metrics, and adversarial testing sets	Adversarial Simulation Suite v2 under development (Q4 2025)
Interoperability with legacy credit and VC systems	Build DID-based bridge adapters and modular plugins	External verifier connector module in Phase III (Scale-Up): 2026
Infrastructure costs for proof issuance and storage	Explore Layer-2 networks and batching models for low-cost operations	Migration to zkRollups or Filecoin layer (exploration in 2026)
No ethics board or governance structure yet operational	Define council charter, voting logic, and accountability model	Publish governance whitepaper and form pilot board by end of 2025

Table 7. Staged Research Roadmap

Staged Research Roadmap:

- **Phase I (Completed)**: Protocol design, rule-based scoring, simulation validation.
- **Phase II (Q2–Q3 2025):** ML engine prototype, real-world pilot with freelancers and verifiers.

- **Phase III (2026)**: Cross-platform integration, infrastructure scaling, formal compliance certification.

12. Conclusion

This paper introduced **Truvry**, a decentralised, behaviour-based trust protocol that enables users to generate and share cryptographically anchored, privacy-respecting proofs of trust. Unlike traditional identity-based verification systems that rely on static credentials or opaque scores, Truvry captures dynamic behavioural signals, allowing trust to be portable and interpretable.

The protocol was validated using simulated datasets and adversarial scenarios. The results demonstrated promising system performance, with low-latency proof issuance, resilience against manipulation, and alignment with regional privacy laws.

Future work will pilot Truvry in participatory-budgeting platforms and citizen assemblies to quantify gains in democratic legitimacy and procedural fairness. Success in these settings would confirm that behaviour-based, portable trust proofs can serve as foundational infrastructure for democratic decision-making in an AI-mediated world.

This work contributes the following:

- **A novel architecture** for portable, behaviour-based trust proofs, anchored on blockchain and compatible with decentralised identifiers (DIDs).
- **An audit-friendly, privacy-preserving framework**, compliant with GDPR, India's DPDP Act, and aligned with W3C verifiable credentials.
- **A six-stage lifecycle model** detailing trust proof generation, revocation, and verifier access.
- **A governance and ethics framework**, grounded in redressal, transparency, and independent oversight mechanisms.

By decoupling trust from identity and reframing it as a composable behavioural asset, Truvry provides a credible alternative for communities and institutions seeking an auditable, behaviour-based trust infrastructure.

About the author

Arpita Pathak is a researcher in human-centred AI, digital autonomy, and algorithmic ethics. Her work focuses on adversarial unlearning, regret-responsive systems, and consent-centric governance models that keep AI aligned with human rights and dignity.

Statements and Declarations

Conflicts of Interest

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency.

Data Availability

This study uses only synthetic, non-identifying data. To mitigate misuse and gaming risks of the trust-scoring protocol, the datasets and prototype scripts are not publicly available; the article provides sufficient methodological detail for assessment, and no personal data were used.

Impact Statement for Policymakers

Truvry provides governments, platform operators, and civic-tech builders with a privacy-preserving way to verify behaviour without harvesting personal identifiers. Portable proofs lower onboarding barriers for excluded workers, curb misinformation by anchoring provenance, and increase auditability for algorithmic public-policy pipelines. Because each proof is user-issued, cryptographically signed, hash-anchored (zk-ready), and revocable, regulators can adopt the protocol without expanding surveillance powers, meeting GDPR data-minimisation duties and India’s DPDP consent principles. By making trust a transferable public good, Truvry advances inclusive digital participation and more transparent, evidence-centred democratic decision-making.

Author Contributions

Arpita Pathak was responsible for the conceptualisation, research, and writing of this manuscript.

Acknowledgements

The authors gratefully acknowledge the contributions of the open-source communities and infrastructure providers whose tools were used in the implementation and validation of Truvry. This includes **Pinata** (IPFS pinning), **MetaMask** (wallet integration), **GitHub Pages** (hosting test proofs) and **Node.js/Express** (backend simulation). While no external reviewers formally contributed to the research, internal testing was supported by simulation volunteers and anonymized data analysis tools. The author also appreciates informal contributions and feedback from internal simulation reviewers and anonymous community testers.

References

1. ^*European Commission (2024). "Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 Amending Regulation (EU) No 910/2014 as Regards Establishing the European Digital Identity Framework." Off J Eur Union. L 2024/1183 (30 April 2024). https://eur-lex.europa.eu/eli/reg/2024/1183/oj/.*
2. ^*Resnick P, Zeckhauser R, Swanson J, Lockwood K (2006). "The Value of Reputation on eBay: A Controlled Experiment." Exp Econ.* **9***(2):79101. doi:10.1007/s10683-006-4309-2.*
3. ^*Slee T (2017). Whats Yours Is Mine: Against the Sharing Economy. 2nd ed. New York: OR Books. ISBN 9781682191200.*
4. ^*uPort Team (2022). "uPort Protocol Specs." GitHub repository. https://github.com/uport-project/specs.*
5. ^*BrightID Team (2023). "BrightID Proof of Uniqueness." BrightID. https://www.brightid.org/.*
6. ^*Lens Protocol Team (2023). "Lens Social Protocol Documentation." Lens Protocol. https://lens.xyz/docs/protocol.*
7. ^*SourceCred Team (2022). "SourceCred A Tool for Communities to Measure and Reward Value Creation." SourceCred. https://sourcecred.io/.*
8. ^*Coordinape Team (2023). "Coordinape Tools to Recognize and Collaborate in Onchain Networks of Trust." Coordinape. https://coordinape.com/.*

9. ^*International Labour Office (2021). World Employment and Social Outlook 2021: The Role of Digital Labour Platforms in Transforming the World of Work. Geneva: ILO. https://www.ilo.org/publications/flagship-reports/role-digital-labour-platforms-transforming-world-work/.*
10. ^*World Wide Web Consortium (W3C) (2022). "Decentralized Identifiers (DIDs) v1.0: Core Architecture, Data Model, and Representations." W3C Recommendation. https://www.w3.org/TR/2022/REC-did-core-20220719/.*
11. ^*Ben-Sasson E, Chiesa A, Genkin D, Tromer E, Virza M (2013). "SNARKs for C: Verifying Program Executions Succinctly and in Zero Knowledge." In: Canetti R, Garay JA, editors. Advances in Cryptology CRYPTO 2013. Vol. 8043. Springer. p. 90108. doi:10.1007/978-3-642-40084-1_6.*
12. ^*Narula N, Vasquez W, Virza M (2018). "zkLedger: Privacy-Preserving Auditing for Distributed Ledgers." In: Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). Renton, WA: USENIX Association. p. 6580. https://www.usenix.org/system/files/conference/nsdi18/nsdi18-narula.pdf.*
13. ^*GitHub (2023). "What Is GitHub Pages?" GitHub Docs. https://docs.github.com/en/pages/getting-started-with-github-pages/what-is-github-pages.*
14. ^*Cabral L, Hortasu A (2004). "The Dynamics of Seller Reputation: Theory and Evidence from eBay." NBER Working Paper No. 10363. doi:10.3386/w10363.*
15. ^*European Union (2016). "Regulation (EU) 2016/679 (General Data Protection Regulation)." Off J Eur Union. OJ L 119:188. https://data.europa.eu/eli/reg/2016/679/oj.*
16. ^*Government of India (2023). "The Digital Personal Data Protection Act, 2023 (No. 22 of 2023)." The Gazette of India. 11 August 2023. https://egazette.gov.in/WriteReadData/2023/248045.pdf.*
17. ^*Pinata Team (2023). "Pinata: Media Pinning and IPFS Gateway Service." Pinata. https://pinata.cloud/.*
18. ^*Organisation for Economic Co-operation and Development (OECD) (2019). "OECD Principles on Artificial Intelligence (OECD Council Recommendation on AI, 22 May 2019)." OECD. https://oecd.ai/en/ai-principles.*

Declarations

Funding: No specific funding was received for this work.
Potential competing interests: No potential competing interests to declare.