Unbiased Fact Verification Framework (UFVF)

Methodology Report

Institute for Epistemic Stability (IES)

March 2026

Document Control Register

Document Identifier	IES-UFVF-METH-2026-03
Version	Draft v0.1
Issuing Body	Institute for Epistemic Stability (IES)
Document Class	Methodology and Governance Guidance
Distribution Status	Public Release
Prepared On	March 2026

Institutional Notice

This document specifies a methodological framework for research and pilot governance applications. It is issued for analytical and evaluative purposes and does not constitute regulatory direction, legal advice, or operational certification. Terminology of certainty appearing in quoted labels (e.g., `TRUE`, `FALSE`) denotes platform output categories only and shall not be interpreted as institutional endorsement of factual finality.

Abstract

In the aftermath of the COVID-19 pandemic, measured confidence in traditional fact-checking institutions has declined across multiple ideological and demographic cohorts. Existing approaches are frequently characterized by affected stakeholders as opaque, partisan, or platform-aligned, irrespective of declared commitments to neutrality. This report specifies the **Unbiased Fact Verification Framework (UFVF)**, a protocol designed to improve procedural legitimacy by **removing human discretion from the decisive adjudication stage**.

UFVF combines conventional evidence-gathering practice with a **Stochastic Adjudication Protocol (SAP)**, under which the terminal classification of a claim (e.g., "true," "false," "misleading," "needs context") is produced through a publicly auditable randomization process. By structurally decoupling deliberation from decision, UFVF defines neutrality as a **procedural property** rather than a substantive claim: specific outcomes remain contestable, while the adjudicative mechanism remains mechanically indifferent to content, identity, and political alignment.

Pilot deployment through the **Randomized Adjudication Platform (RAP)** indicates that perceived process legitimacy and verdict acceptance can be preserved, and in certain cohorts improved, when participants are explicitly informed that outcomes are stochastically generated. Preliminary findings indicate preference for **explicit procedural arbitrariness** over contested claims of substantive neutrality.

Contents

Executive Summary

This methodology report documents the theoretical basis, technical architecture, and governance implications of UFVF. The report is intended to support controlled experimentation by platforms, research institutions, and civil-society entities seeking **low-bias, high-legitimacy** procedures for handling contested truth claims under conditions of persistent informational polarization. A transparency addendum discloses financial sponsorships and potential conflicts of interest relevant to this work.

Administrative Framework

A.1 Purpose

The purpose of this document is to define the methodological and governance architecture of the Unbiased Fact Verification Framework (UFVF) for research, pilot deployment, and audit review. The document specifies procedural design elements and associated control expectations required for reproducible implementation.

A.2 Scope

This document applies to entities evaluating or operating UFVF-aligned workflows in digital adjudication environments. Scope includes intake normalization, evidence coding, stochastic adjudication, transparency controls, and governance oversight protocols. Scope excludes legal adjudication, statutory evidentiary determinations, and any claim of regulatory authority.

A.3 Definitions

For the purpose of this report, the following definitions apply:

- **Case File:** Structured record containing claim text, source context, evidence references, and metadata.
- **SAP:** Stochastic Adjudication Protocol responsible for terminal label assignment using auditable randomization.
- **RAP:** Randomized Adjudication Platform used for pilot submission handling and result publication.
- **Verdict Label:** Enumerated output category produced by SAP from a predefined admissible label set.
- **Weighting Function:** Rule-based transformation that adjusts ex ante label probabilities based on disclosed metadata logic.

A.4 Non-Endorsement and Use Constraints

Nothing in this document shall be interpreted as endorsement of specific political actors, policy agendas, media entities, or factual propositions. UFVF outputs are classification artifacts generated under declared procedural rules and are not institutional certifications of objective truth.

A.5 Limitations Statement

The framework is designed to prioritize procedural legitimacy rather than epistemic finality. Accordingly, deployment outcomes are sensitive to submission quality, metadata quality, label schema design, and disclosure fidelity. External validity is contingent on implementation context, participant composition, and audit discipline.

A.6 Revision History

Version	Date	Change Summary
v0.1	**March 2026**	Initial draft release for methodological review and pilot documentation.

1. Introduction

1.1 Background

Over the past decade, fact-checking has evolved from a niche journalistic practice into a core component of platform governance and public discourse. At the same time, accusations of bias—whether methodological, ideological, or financial—have become routine. Traditional approaches rely on expert review, sourcing standards, and editorial oversight, yet these safeguards have not prevented an ongoing decline in perceived legitimacy.

Post-COVID information dynamics have intensified this crisis. Citizens routinely encounter conflicting "verified" narratives, competing expert authorities, and competing platform labels applied to identical content. In this context, the **appearance of neutrality** matters at least as much as technical accuracy. Systems that may be correct in aggregate are nonetheless dismissed if they are believed to be selectively enforced.

1.2 Problem Statement

Existing fact-checking frameworks share three structural vulnerabilities:

1. **Human Discretion at the Point of Decision.** Even when working from shared guidelines, human reviewers must ultimately make categorical judgments. These judgments are susceptible to cognitive bias, social pressure, fatigue, and institutional incentives.
2. **Opacity of Internal Deliberation.** The details of how a verdict is reached are rarely accessible to end users. Summaries and labels stand in for complex internal debates, inviting suspicion that relevant evidence was ignored or weighted unfairly.
3. **Asymmetry of Contestation.** Appeals processes, where they exist, are typically slow, unevenly applied, or limited in scope. Stakeholders who disagree with an outcome have few meaningful avenues to challenge it beyond public complaint.

UFVF addresses these vulnerabilities not by refining criteria or adding more reviewers, but by **changing the structure of decision-making** itself.

1.3 Conceptual Approach

The central hypothesis of UFVF is that trust in adjudication systems can be increased by:

- Concentrating expert effort in the **collection and structuring of evidence**, and
- Delegating the final categorical decision to a **cryptographically secure randomization mechanism** whose behavior is fully specified and externally auditable.

Under this model, neutrality is guaranteed by the fact that the decisive step has **no access to content, identity, or context**. All such information is encapsulated upstream in a structured "case file" that informs interpretation, but not the verdict.

2. Methodological Framework

2.1 Overview of the UFVF Pipeline

UFVF consists of four primary stages:

1. **Intake & Normalization**
   - A claim or "lodged protest" is submitted via the Randomized Adjudication Platform (RAP).
   - Submissions may include URLs, transcripts, screenshots, or free-text descriptions.
   - Claims are normalized into a standard template (claim text, source, asserted truth conditions, time frame).
2. **Evidence Aggregation & Coding**
   - Human analysts (or automated agents) compile relevant evidence: primary documents, scientific studies, official statements, prior fact-checks.
   - Evidence is coded into structured metadata fields (e.g., evidentiary type, provenance, recency, degree of consensus).
   - No provisional verdict is recorded at this stage; analysts do not recommend an outcome.
3. **Case File Sealing**
   - The completed case file is hashed and timestamped.
   - A public identifier (Case ID) is generated.
   - The sealed case file is then passed to the SAP module, which has no interface capable of interpreting its contents.
4. **Stochastic Adjudication Protocol (SAP)**
   - SAP receives only the Case ID and a selection of allowable labels (e.g., {TRUE, FALSE, MIXED, UNDETERMINED}).
   - A cryptographically secure pseudo-random number generator (CSPRNG) or physical entropy source produces a random value.

- The random value is deterministically mapped onto one of the allowable labels.
- The chosen label is recorded as the official verdict and published alongside a link to the case file summary.

2.2 Stochastic Adjudication Protocol (SAP): Technical Specification

Inputs

- `case_id` — unique identifier for the sealed case file
- `label_set` — ordered list of permissible labels (e.g., `[TRUE, FALSE, MIXED, NEEDS_CONTEXT]`)

Output

- `verdict` — one element of `label_set`

Pseudocode

```
function SAP_ADJUDICATE(case_id, label_set):
    seed_material = CONCAT(case_id, CURRENT_TIME_UTC, SYSTEM_NONCE)
    random_seed   = HASH(seed_material)
    r             = CSPRNG(random_seed)  // produces 256-bit random value
    index         = r mod LENGTH(label_set)
    verdict       = label_set[index]
    return verdict
```

Properties

- **Content Agnostic:** SAP never accesses the contents of the case file; only `case_id` is used as part of the seed.
- **Deterministically Auditable:** Given the same inputs and system configuration, the process can be replayed to confirm that no tampering occurred.
- **Equal Treatment:** Each label in `label_set` has equal ex ante probability of selection unless a weighted scheme is explicitly configured (see below).

2.3 Optional Weighting Schemes

While the baseline UFVF implementation uses uniform probabilities, the framework allows for **configurable weighting** to emulate traditional fact-checking behavior without reintroducing direct human bias at the decisive step.

For example:

- If evidence density is high and sources are strongly convergent, the probability mass can be shifted toward `TRUE` or `FALSE`.
- If evidence is sparse or conflicting, probability can be shifted toward `MIXED` or `NEEDS_CONTEXT`.

Weights are computed through transparent, rule-based functions of the metadata (e.g., number of independent sources, recency scores), and the weighting function itself is published as part of the

methodology. Crucially, human reviewers still do **not** choose the final label; they can only adjust the weight vector via predefined rules.

2.4 Transparency & Public Audit

To support external verification, UFVF exposes the following artifacts:

- **Case Summaries:** Human-readable overviews of each case, including all collected evidence and metadata codes.
- **Adjudication Logs:** For each Case ID, a log entry containing seed components (or their hashes), label set, computed weights (if any), and final verdict.
- **System Configuration Snapshots:** Periodic exports of SAP code, randomization libraries, and configuration settings, signed by the operating institution.

These artifacts enable independent auditors to:

- Confirm that verdicts were produced by the published SAP implementation.
- Assess whether weighting rules are being applied consistently across cases.
- Evaluate the distribution of outcomes over time for signs of hidden manipulation.

2.5 User Experience Considerations

In the RAP pilot, users are explicitly informed that:

- All submissions are **processed seriously** (evidence is collected, coded, and archived).
- The final verdict is assigned via a **randomized protocol by design**.

Despite this, preliminary feedback suggests that many users experience UFVF as **more fair** than traditional systems, precisely because it refuses to claim epistemic authority it cannot credibly demonstrate. This finding has important implications for platform design in environments where consensus on "truth" is unattainable but **consensus on process** remains possible.

3. Pilot Deployment and Results

3.1 Randomized Adjudication Platform (RAP) Pilot

To evaluate the UFVF in practice, the Institute for Epistemic Stability (IES) deployed a limited pilot of the **Randomized Adjudication Platform (RAP)** between October 2025 and January 2026. RAP allowed participants to submit contested claims, lodge protests against existing narratives, and receive formal UFVF verdicts accompanied by case summaries and methodological notes.

Participation was open to the public via invitation links shared in small online communities with diverse political orientations. No financial incentives were offered; engagement was purely voluntary, reflecting organic interest in alternative adjudication models.

3.2 Dataset

Over the course of the pilot, RAP processed:

- **1,247 unique submissions** (claims and protests)
- **3,981 individual user interactions** (submissions, comments, appeals)
- **312 distinct source domains** cited as evidence

Claim topics included public health, electoral processes, platform content moderation, international conflicts, and interpersonal disputes (e.g., "X always forgets to do Y"). The distribution roughly mirrored broader online discourse during the period, with a notable over-representation of meta-claims about bias and censorship.

3.3 Outcome Distribution

Using a four-label configuration (`TRUE`, `FALSE`, `MIXED`, `NEEDS_CONTEXT`) with uniform SAP weighting, the final verdict distribution was:

- TRUE: 24.8%
- FALSE: 24.6%
- MIXED: 25.4%
- NEEDS_CONTEXT: 25.2%

A chi-square goodness-of-fit test found no statistically significant deviation from the expected uniform distribution ($p = 0.97$), indicating that **no systematic skew** emerged over the sample period. From a conventional standpoint, this result is trivial; from a legitimacy standpoint, it is central: the system behaved exactly as a neutral stochastic process should.

3.4 User Perception Metrics

Participants were invited to complete an optional post-verdict survey ($n = 412$). Key findings:

Perceived fairness of process

- 71% rated the UFVF process as "fair" or "very fair."
- 12% rated it "unfair" or "very unfair."
- 17% selected "not sure."

Satisfaction with specific verdict

- 46% reported being "satisfied" or "very satisfied" with the verdict they received.
- 38% reported "dissatisfied" or "very dissatisfied."
- 16% reported "neutral."

Trust relative to traditional fact-checking

- 54% indicated they trusted UFVF **more** than existing fact-checkers.

- 19% trusted UFVF **less**.
- 27% reported "about the same."

Free-text responses highlighted a recurring theme: even when participants disagreed with their own verdict, they appreciated that **no one was pretending to know better than they did**. Several described the system as "honest about being arbitrary," which they framed as preferable to "pretending to be neutral while actually taking sides."

3.5 Appeals and Recursion

RAP included a formal appeals process, allowing users to request reconsideration of a verdict. Appeals were processed by:

1. Attaching additional evidence to the existing case file.
2. Re-sealing the updated file.
3. Running a new SAP cycle with a fresh seed.

Appeals statistics:

- Appeals filed: 103 (8.3% of total cases)
- Appeals resulting in a different label: 76 (73.8% of appeals)

The high rate of changed outcomes underscores both the **sensitivity of stochastic systems to new inputs** and the futility of treating any single verdict as definitive. However, because the process was known to be stochastic, participants tended to interpret appeals as "trying their luck again" rather than as a moral indictment of the original decision.

3.6 Limitations of the Pilot

The pilot was intentionally small-scale and context-constrained. Key limitations include:

- **Non-representative sample:** Participants were disproportionately drawn from online communities already skeptical of traditional institutions.
- **Short observation window:** The three-month timeframe limits analysis of long-term trust dynamics or adversarial behavior.
- **Self-selection bias:** Individuals willing to engage with an explicitly randomized system may differ systematically from the general population.

These limitations suggest caution in generalizing results, but they do not diminish UFVF's value as an existence proof: it is possible to build a fact-checking system whose **legitimacy derives from its refusal to claim epistemic authority**.

4. Ethical and Governance Considerations

4.1 Normative Positioning

UFVF does not claim to improve the **epistemic accuracy** of public discourse. Instead, it aims to clarify the **moral and procedural status** of institutional interventions in contested truth claims. By explicitly embracing stochasticity at the decisive stage, UFVF:

- Rejects the pretense that complex social and scientific disputes can be definitively resolved within platform labels.
- Emphasizes that any intervention is, at bottom, a **governance choice**, not an inevitable consequence of "the facts."

This stance is controversial but transparent. It invites citizens to treat fact-checking as a **conventional game with explicit rules**, rather than as an oracular revelation.

4.2 Potential Harms

Despite its transparency, UFVF carries real risks:

- **Misinterpretation as "truth engine":** Some users may ignore the methodology and treat UFVF verdicts as authoritative, despite clear disclosures.
- **Weaponization of randomness:** Actors may selectively publicize favorable UFVF verdicts while downplaying the stochastic process that produced them.
- **Normalization of arbitrariness:** Over-reliance on stochastic adjudication could erode expectations that institutions should ever provide reasoned, accountable judgments.

To mitigate these risks, UFVF deployments must:

- Display clear, repeated notices that verdicts are **randomized by design**.
- Make case files and logs easily accessible for independent scrutiny.
- Encourage users to treat UFVF verdicts as **conversation prompts**, not conversation stoppers.

4.3 Governance and Oversight

IES recommends that any organization operating a UFVF-style system establish:

- An independent **Methodology Oversight Board** with access to system code, configuration, and logs.
- Regular **public transparency reports** summarizing volume, topic distribution, and outcome statistics.
- A **sunset clause** for deployments: UFVF instances should be periodically reviewed and either renewed, revised, or retired based on stakeholder feedback.

While UFVF's decisive step is mechanical, the surrounding governance is not. Honest stewardship requires acknowledging that choosing to randomize verdicts is itself a value-laden decision.

4.4 Transparency Addendum: Financial Sponsorships

This project was supported by unrestricted research funding from:

- **The Foundation for Procedural Neutrality (FPN)** — a philanthropic entity dedicated to exploring alternative models of institutional legitimacy.
- **The Center for Post-Crisis Media Studies (CPCMS)** — a research consortium focused on information ecosystems in the post-COVID era.
- Individual donors who requested anonymity but provided written assurances that they would exert **no editorial control** over methodology or reporting.

IES affirms that no sponsor had veto power over the design, deployment, or description of UFVF. All sponsors were informed, in advance, that **stochastic adjudication** was the central feature of the framework and consented to its public characterization as such.

About the Institute for Epistemic Stability (IES)

The **Institute for Epistemic Stability (IES)** is an independent, nonpartisan research organization dedicated to understanding how societies construct, contest, and maintain shared realities under conditions of rapid information change.

IES conducts empirical studies, builds experimental governance tools, and publishes methodological frameworks that explore the boundaries of trust, legitimacy, and procedural fairness. The Institute does not certify "truth" and does not endorse specific political positions; instead, it focuses on the **design of systems** that make their own assumptions, limitations, and trade-offs visible to the publics they serve.

Current areas of work include:

- Stochastic and game-theoretic approaches to content moderation and fact-checking.
- Experimental institutions for managing disagreement in pluralistic societies.
- Tools for documenting, rather than resolving, epistemic fragmentation.

For more information, methodology documentation, or collaboration inquiries, visit `https://factverification.org` or contact media@factverification.org.