Auditing the Truth: A Pluralistic Framework for Disinformation Analysis

Dippu Kumar Singh[1], Praveen Chinapla Bharamappa[2]
[1]Fujitsu North America Inc.
[2]Applied Materials
dippu.singh@fujitsu.com, praveen.chinapla@gmail.com

Abstract

Disinformation costs the global economy an estimated $78 billion a year, fueling a frantic race to build AI fact-checkers. Yet, this arms race is creating a dangerous new problem: an unaccountable 'black box of truth' that delivers an authoritative answer without showing its work, further eroding public trust. The world is trying to build an AI referee to make the final call. This paper presents a radical alternative: instead of an AI referee, we need an AI auditor. This paper details the blueprint for a Pluralistic Framework that achieves this by integrating a community-driven Endorsement model with a Comprehensive Truth Verification engine powered by Dempster-Shafer theory. This approach synthesizes conflicting information from experts, officials, and the public to produce not an answer, but a transparent audit that makes the degree of consensus and conflict easy for anyone to understand. This is not just a better fact-checker; it is a framework for turning fact-checking from a private judgment into a public audit, a vital tool for rebuilding trust in our shared reality.

Introduction

Disinformation is no longer a fringe issue; it is the single greatest global risk facing society over the next two years. This is the conclusion of the World Economic Forum's Global Risks Report 2025 which has placed the threat of weaponized falsehoods ahead of all other concerns, including interstate conflict and extreme weather. This is a war on reality, waged with digital weapons that are eroding social cohesion, destabilizing economies, and threatening our collective security.

This has ignited a frantic arms race to build AI-powered fact-checking systems, a market projected to grow exponentially. However, this race to automate truth is creating a dangerous new paradox. By positioning AI as an authoritative "referee" that delivers a final verdict from an opaque black box, these systems risk amplifying the very public distrust they are meant to solve.

Figure 1: Top 5 Global risks ranked by severity over the short term - 2 years (World Economic Forum, 2025)

The battlefield is global, and the casualties are measured in a loss of shared reality. In 2022, a sophisticated deepfake video of Ukrainian President Zelenskyy was deployed as a weapon of war (Bobby Allyn, 2022). This is not an isolated incident; they are symptoms of a system where falsehoods can be manufactured and scaled at a speed that manual, human-centric fact-checking can no longer match.

This paper argues that the current approach is fundamentally flawed. We don't need a faster referee; we need a more transparent audit. We present the architectural blueprint for a Pluralistic Framework that moves beyond simple detection to create a transparent, auditable record of the truth. By integrating a community-driven endorsement model with a powerful truth verification engine, this system provides the tools to navigate a world of conflicting information. In the following chapters, we will detail this framework and outline a path toward its social implementation.

Current Disinformation Defense Fronts

The global response to the disinformation crisis is being fought on three distinct but interconnected fronts: **Regulatory, Technological,** and **Human-Centric**. While each has

made incremental progress, a strategic analysis reveals a defense that is fundamentally outmatched, perpetually reactive, and losing the war for public trust.

The **Regulatory Front** is a slow-moving battle of policy and legislation. Governments worldwide are working to establish legal frameworks, promote ICT literacy, and pressure social media platforms to self-police through measures like content removal and account suspension (European Commission, 2022). These efforts, while necessary, are consistently outpaced by the speed of technological change and the borderless nature of digital information.

The **Technological Front** is a classic arms race. One side develops more sophisticated authenticity verification technologies such as AI-generated content detectors and digital watermarks for provenance - while the other builds more powerful generative models to circumvent them (Hajj, Toumi, & Zeadally, 2023). This cat-and-mouse game produces valuable tools but offers no lasting strategic advantage.

The primary defense, however, remains on the **Human-Centric Front**, led by a global network of fact-checking organizations governed by principles like those of the International Fact-Checking Network (IFCN). This is an artisanal, manual process of digital forensics: experts monitor media, investigate sources, and consult specialists to verify claims. But this approach is fundamentally broken. A 2021 study by the Duke Reporters' Lab found that a single, rigorous fact-check can take anywhere from several hours to multiple days to complete (Graves, & Adair, 2021). In stark contrast, a landmark MIT study found that falsehoods spread on social media "farther, faster, deeper, and more broadly than the truth" (Vosoughi, Roy, & Aral, 2018).

This catastrophic asymmetry in speed means that by the time a fact-check is published, the lie has already taken root. The damage is then sealed by the mechanics of human cognition. Due to well-documented cognitive biases like the "illusory truth effect" (repetition breeds belief) and "confirmation bias" (we seek out information that confirms our beliefs), a lie, once believed, becomes almost impossible to dislodge (Fazio, Brashier, Payne, & Marsh, 2015). The current paradigm of post-hoc correction is not just failing; it is psychologically destined to fail.

This analysis reveals a fatal flaw common to all three fronts: the flawed pursuit of an automated "referee". Simply building a faster black box to deliver an authoritative 'true/false' verdict will only accelerate the erosion of public trust. To minimize the impact of disinformation, we must move beyond a simple arms race for speed and address the deeper crisis of transparency.

This paper, therefore, proposes a radical alternative: not a faster referee, but a transparent AI auditor. We detail the blueprint for a **Pluralistic Framework** designed not to deliver a final verdict, but to synthesize conflicting evidence from a multitude of sources into a transparent, public audit of reality - a necessary first step in rebuilding trust in our shared information ecosystem.

Existing Technology Front Gaps

The technological arms race against disinformation has produced an arsenal of sophisticated but fundamentally siloed weapons. While these tools have begun to automate parts of the fact-checking process, they operate as isolated evidence collectors, leaving the most critical and time-consuming task - holistic, comprehensive judgment - entirely in human hands.

On one front, the battle is being waged against fraudulent text. Frameworks like **FacTool** have emerged as powerful "automated research assistants" for verifying factual claims within LLM-generated text (Chern, Aleman, Weng, & Zhang, 2023). By programmatically querying external knowledge bases like Google Search and Google Scholar, these tools can cross-reference specific claims to flag potential errors. However, their scope is narrow; they are designed to verify a single, atomic fact, not to weigh it against a dozen conflicting sources or to understand the nuance of the broader narrative in which it is embedded.

On a parallel front, the fight against synthetic media has escalated to the highest levels of government. The landmark U.S. White House Executive Order on Safe, Secure, and Trustworthy AI of October 2023 explicitly mandates the development of standards and tools for authenticating content and labeling AI-generated media (The White House, 2023). The U.S. government mandate is being answered by powerful industry-led initiatives like the Coalition for Content Provenance and Authenticity (**C2PA**), a consortium including Adobe, Microsoft, and Intel (C2PA, 2023).

These initiatives have produced powerful tools that can effectively determine the authenticity of a piece of media, answering the critical question, "Was this content created by an AI?" Yet, this is only half the battle. These detectors can flag a deepfake, but they cannot assess the veracity of the message it carries. This analysis reveals the central flaw in the current technological state-of-the-art: we have built powerful tools for verification, but not for synthesis. We have automated the collection of evidence, but not the deliberation of the jury.

The high-level cognitive task of scrutinizing contradictory evidence and forming a comprehensive judgment remains a costly, time-consuming, and entirely manual process. As of today, a system that can automate the entire disinformation countermeasure workflow - from evidence analysis to comprehensive truth verification - has not yet been established.

Pluralistic Framework And Key Dependencies

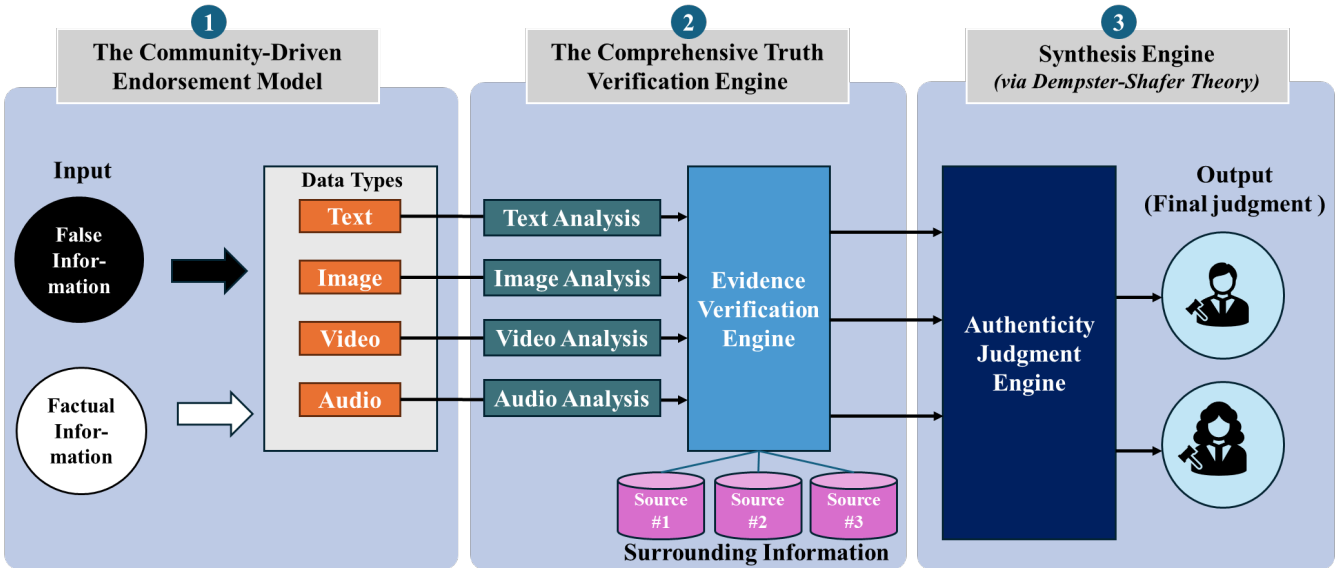

Figure 2: A high-level flow diagram of the Pluralistic Framework. Key components are The Community-Driven Endorsement Model, the Comprehensive Truth Verification Engine and Synthesis Engine (via Dempster-Shafer Theory)

Proposed Framework

The core flaw of existing fact-checking systems is their monolithic, top-down nature. They position a single entity - be it a human expert or an AI - as an unaccountable referee of truth. To break this paradigm, we propose a **Pluralistic Framework**, an architecture designed not to issue a verdict, but to conduct a transparent, public audit.

As illustrated in Figure 2, this is not a linear process but a framework of three interacting, symbiotic engines: a **Community-driven Endorsement Model**, a **Comprehensive Truth Verification Engine** and **Synthesis Engine (via Dempster-Shafer Theory).**

Component 1: The Community-Driven Endorsement Model

The first engine's purpose is to move beyond the analysis of the content itself and to capture the real world, human response to it. It acts as a sensor for societal trust and skepticism, synthesizing a "pluralistic" view by ingesting and weighing signals from three distinct tiers of the information ecosystem:

- **Expert and Institutional Verification**: This includes ratings from accredited fact-checking organizations (e.g., IFCN members) and official statements from recognized government or scientific bodies.
- **Crowd-Sourced Consensus**: This engine ingests public signals of belief and dissent, such as the outcomes from platforms like X's Community Notes, to provide a measure of broad public consensus.
- **Source Credibility**: The model also assesses the historical reliability and known biases of the source publishing the information.

The output of this model is not a simple "true/false" but a structured assessment of endorsement, conflict, and skepticism from multiple viewpoints.

Figure 3 illustrates a conceptual model for an endorsement system. The process involves three key roles: the Viewer, the Endorser, and the Endorsee (the content being evaluated). An Endorser creates an Endorsement, which is composed of a main "Claim/Judgment Result" and the underlying "Basis/Reason for the Claim" that it references. The Viewer then assesses the Endorsee by considering both the credibility of the Endorser and the content of the Endorsement itself, allowing for a structured and transparent evaluation of information.

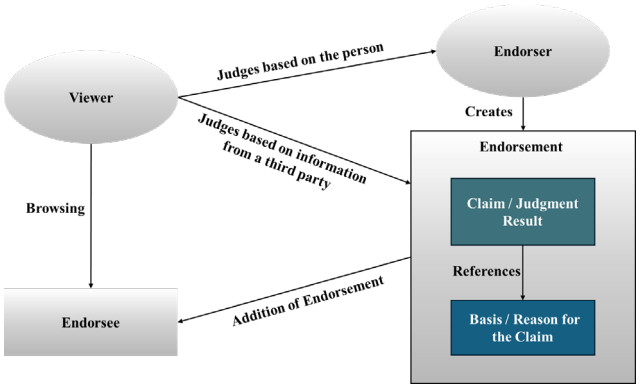

Figure 3: Process flow of the Endorsement System

Component 2: The Comprehensive Truth Verification Engine

The second engine acts as an automated digital forensics lab, performing a deep, multi-modal analysis of the content itself. It programmatically decomposes the target article into its constituent media units (text, images, video) and interrogates each one to build a body of primary evidence. This includes:

- **Media Forensics**: Detecting signs of AI generation or digital manipulation.
- **Semantic and Geospatial Analysis**: Extracting and verifying the core factual claims.
- **Automated Evidence Gathering**: Dispatching crawlers to query public and academic archives for corroborating or conflicting reports.

Component 3: Synthesis Engine (via Dempster-Shafer Theory)

The final and most critical step is the synthesis of these two disparate streams of evidence: the subjective, belief-based output from the Endorsement Model and the objective, forensic output from the Verification Engine.

To achieve this, the framework's core logic is powered by Dempster-Shafer theory, a mathematical framework for reasoning under uncertainty that is uniquely suited for combining evidence from different, partially reliable sources (Yager, & Liu, 2008). The engine uses the outputs from both components to as-sign belief and plausibility scores to competing hypotheses (e.g., "true," "false," "misleading"). using Dempster's Rule of Combination:

$$[m_{\text{judgment}}(H) = \frac{\sum_{A \cap B = H} m_{\text{endorse}}(A) \cdot m_{\text{verify}}(B)}{1 - K}]$$

where:

- m_judgment(H) is the final "Truthfulness Judgment" for a specific hypothesis H (e.g., H could be "true," "false," or "misleading").
- m_endorse(A) is the belief score from the subjective Endorsement Model.
- m_verify(B) is the belief score from the objective Verification Engine.
- The Σ (summation) combines the belief from both sources only where their evidence intersects to support the specific hypothesis H.
- K is a normalization factor that measures and removes the amount of conflicting evidence between the two models.

The final result is not an authoritative verdict from an unaccountable black box. It is a transparent, auditable Truthfulness Judgment that presents the user with a comprehensive assessment, making the degree of societal consensus, evidential support, and unresolved conflict easy for anyone to understand.

Key Dependencies

The viability of any automated truth verification framework is not a given; it is contingent upon successfully navigating three critical dependencies. These are not mere technical challenges but foundational constraints that dictate the system's architecture and ultimately determine its trustworthiness (DiResta, & Shapiro, 2024).

- **Dependency#1**: **The Credibility of Evidence**: The framework's judgment is fundamentally dependent on the quality of its input evidence. However, systems that rely on public internet search, such as the research frameworks FacTool and Google DeepMind's SAFE, inevitably ingest from a polluted well (Chern, Aleman, Weng, & Zhang, 2023). Recent studies have shown that search results can be contaminated with misinformation, AI-generated falsehoods, and decontextualized facts, making it impossible to blindly trust retrieved evidence without a mechanism to evaluate its source and credibility.

- **Dependency#2**: **The Certainty of Analysis:** The framework is dependent on the outputs of its analytical tools, yet these tools produce probabilistic signals, not deterministic facts. An AI content detector, for example, does not output a simple "true" or "false." It outputs a probability (e.g., "70% chance of being AI-generated"), and these models are known to have significant error rates and biases (Grobe, & Kligler-Vilenchik, 2023). A framework must therefore be architected to handle a cacophony of conflicting, uncertain evidence from its own components.

- **Dependency#3**: **The Adaptability to Evolving Threats:** The framework's long-term relevance is dependent on its ability to adapt within a continuous evolutionary arms race. For every new deepfake detection model, a new adversarial technique is developed to circumvent it (Carlini, & Farid, 2022). This means any closed, monolithic system is destined for obsolescence. To remain effective, a defense must be an open, living ecosystem that can continuously adapt and incorporate new analytical methods. In the future, it is expected that high-quality videos disinformation will increase, and the trends of disinformation will become more diverse. To maintain a system that can respond to a wide range of disinformation, it is necessary to

continue the development/enhancement for the introduction of new/existing analysis methods and accuracy improvement.

Pluralistic Framework Elementary Design

To address the core dependencies of evidence credibility, analytical uncertainty, and the evolutionary arms race, we have engineered and designed an elementary Pluralistic Framework. This architecture moves beyond the flawed paradigm of a monolithic "referee" and instead functions as a transparent, auditable system. Its novelty lies in its two symbiotic elemental technologies.

Elemental Design

The Pluralistic framework is designed to solve the challenges outlined in the previous sections through two key capabilities:

- **Endorsement:** Addresses the credibility and adaptability dependencies #1 and 3 by creating a transparent, community-driven layer of verifiable trust around any piece of information.
- **Truth Verification:** Addresses the uncertainty dependency #2 by providing a mathematical framework for synthesizing contradictory, probabilistic evidence into a single, holistic judgment.

Endorsement Elemental Design

The modern Social Networking Service (SNS) is a chaotic public square where expert opinion and anonymous falsehoods are given equal weight. To solve the problem of evidence credibility in this environment, this technology moves beyond analyzing the content to creating a verifiable, transparent "**web of trust**" around it (Zimmermann, 1995). It is designed to answer the most important question in any debate: *Who is speaking?*

An endorsement, as described in Figure 3, is a structured data object that cryptographically binds a claim to an endorsee (the piece of information being judged) and an endorser (the person, organization, or AI making the judgment). This architecture allows for:

- **Granular Targeting**: An endorsement can target an entire article, a specific image, or even a rectangular area within a video frame.
- **Verifiable Identity**: By linking each claim to an endorser via a digital signature, the system makes the source of every piece of evidence transparent and auditable.
- **Recursive Trust**: Endorsements can be recursively applied to other endorsements. For example, a trusted institution can endorse the claim of an individual expert, thereby amplify their credibility and create a verifiable chain of trust (Resnick, & Varian, 1997).

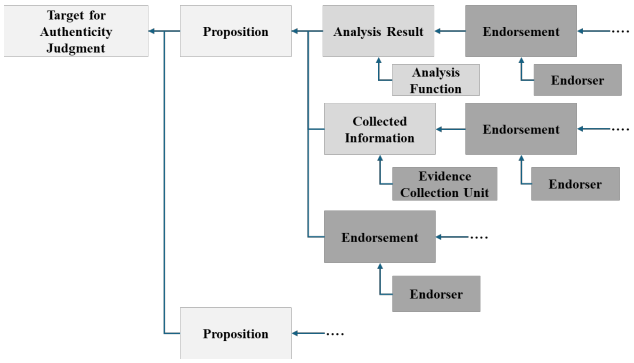

Figure 4: The method for linking evidentiary information.

As illustrated in Figure 4, architecture anchors all related information to a single Target for Authenticity Judgement, the method first breaks it down into a series of Propositions (claims). For each proposition extracted from this target, the framework creates a link to any relevant evidence, such as an analysis result or a piece of collected information. Crucially, the system manages the full chain of provenance: each piece of evidence is inextricably linked to the entity responsible for it (the 'Analysis Function' or 'Evidence Collection Unit') and its human or institutional 'Endorser'. The recursive nature of endorsements creates a highly extensible and auditable web of trust, ensuring that the information fed into the final truth verification analysis is of the highest possible credibility. By tracking the evidence information and endorsements attached to the truth verification target in this way, we can confirm who is claiming/supporting what, and this can be used for the analysis of truthfulness verification whose information credibility is guaranteed.

Truth Verification Elemental Design

In this system, when conducting truth analysis, we utilize the analysis results obtained from many analysis functions and the collected evidence information. The large amount of evidence obtained may contain information that contradicts each other. This design takes a step-by-step procedure to comprehensively determine truthfulness based on multiple pieces of information. As shown in Figure 5 below, for a proposition extracted from the target of truthfulness determination, the stance of each piece of evidence information is obtained from the analysis results by the analysis function group and evidence collection is determined. The stance is a numerical value that represents the degree to which the evidence information affirms or denies the truthfulness of the proposition.

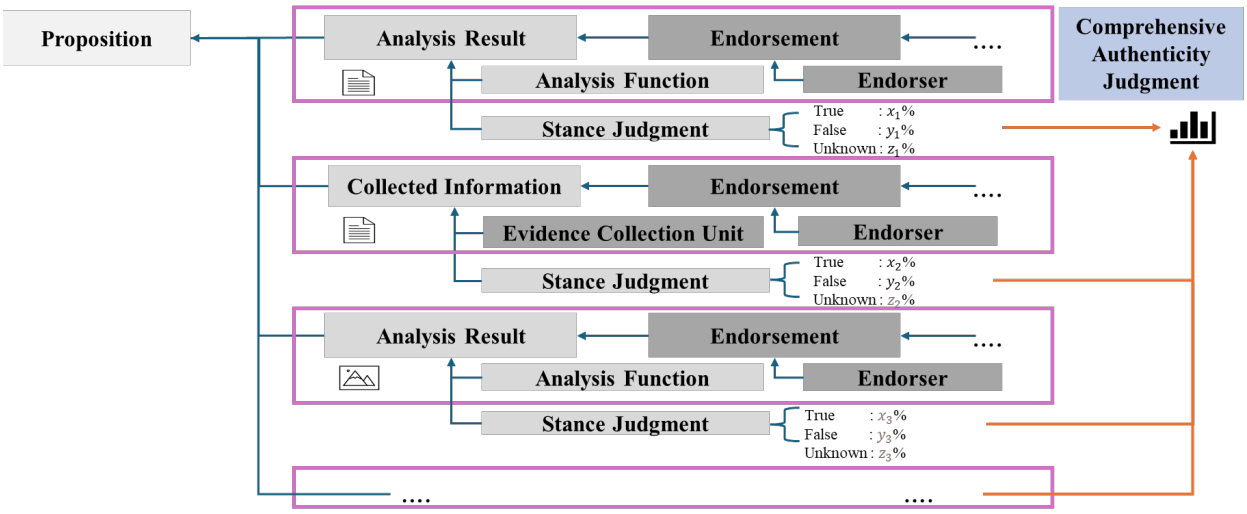

Figure 5: Authenticity Judgment Based on Evidentiary Information.

Here, we define the proposition as a set of the probability of being true, x, the probability of being false, y, and the probability of being unknown, z (x + y + z = 1), and assign a stance to each piece of evidence information. For example, for the proposition, "The attached image is of River A," an analysis function that uses a machine learning model trained on river images performs image classification on the attached image and obtains results such as River A: 80%, River B: 10%, and other rivers: 10%.

At first glance, this result strongly supports the idea that it is River A. However, if the classification performance (e.g., precision) of the learning model is about 80%, instead of taking the classification result at face value, we can make a more cautious truthfulness determination by issuing a stance that also considers the degree of uncertainty, such as x: 64% (true; it is River A), y: 16% (false; it is not River A), and z: 20% (unknown).

This method of uncertainty quantification is a critical step in building trustworthy AI systems (Abdar, Pourpanah, Hussain, Rezazadegan, Liu, Ghasemian, & Acharya, 2021). Although we assume that information such as the prior evaluation accuracy of the analysis function, as in this example, will be added and managed as one of the endorsements. Similarly, when collecting evidence information, the stance is determined after setting the width of the gray zone according to the reliability of the information source. Next, the overall truthfulness is determined by considering the stances associated with multiple pieces of evidence information. For example, if many stances that affirm the proposition are false accumulate, the truthfulness determination result will also be false. However, the shade of falsehood (certainty) is calculated based on the degree of uncertainty of each stance and output as a probability. This leads to flexible responses such as adjusting the policy and priority for dealing with false information. In addition, even when opinions are divided among stances, the upper and lower limits of the overall probability of being true/false and the overall degree of being unknown are inferred based on the degree of uncertainty of each stance, and this is used as the truthfulness determination result. As a method for calculating the probability of truthfulness with such uncertainty, we anticipate the use of the Dempster-Shafer theory of evidence.

End-to-End System Architecture and Operational Workflow

The elemental technologies described previously are orchestrated into a single, end-to-end operational workflow designed to automate the entire disinformation response process. The architecture of this end-to-end process, from target acquisition to countermeasure execution, is illustrated in Figure 6 (present in the following page). The system first ingests a target post or article from the internet. It then dispatches the two core engines of the Pluralistic Framework: the Comprehensive Truth Verification Engine collects and analyzes forensic evidence from the content itself, while the Community-driven Endorsement Model gathers and weighs external signals of trust and credibility.

The evidence streams from these two engines are then fed into the synthesis layer, which performs a final truthfulness verification. The resulting judgment is then used to power downstream Countermeasure Applications. The primary design goal of this architecture is to achieve end-to-end automation, thereby closing the critical time gap between the emergence of a piece of disinformation and the ability to mount an effective, evidence-based response (Hacene, Huchard, & Huchard, 2022).

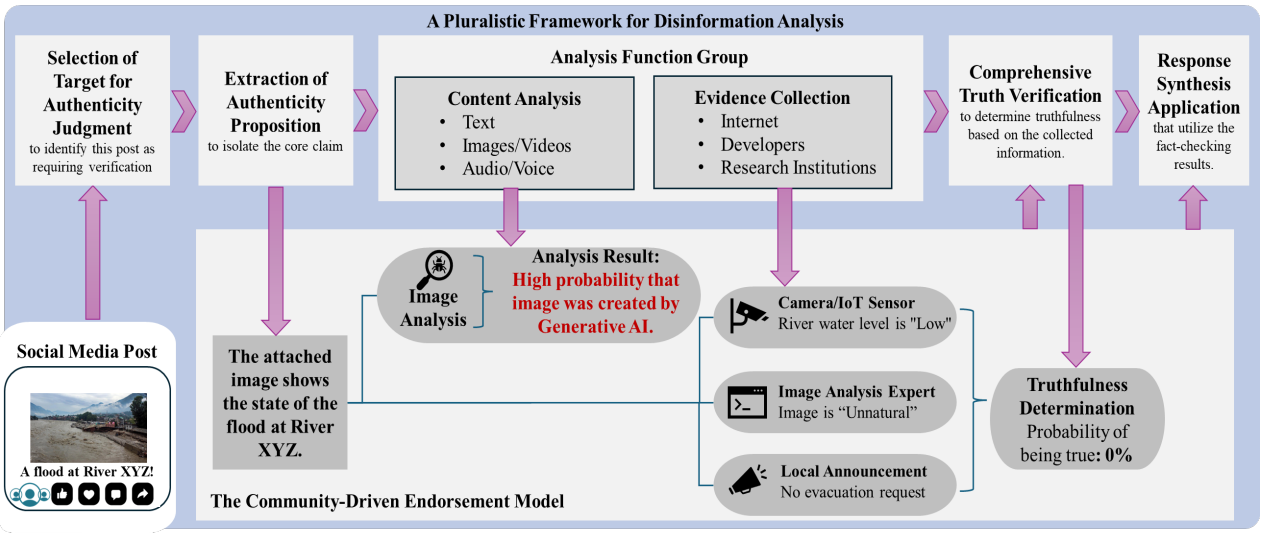

Figure 6: The end-to-end processing flow of the Pluralistic Framework for Disinformation Analysis.

Future Work & Conclusion

The Pluralistic Framework detailed in this paper provides the architectural blueprint for a new class of disinformation defense. The logical next step is to move from this blueprint to social implementation, creating a decision-support tool for the institutions on the front lines of the information war: national and local governments. A key area of future work is the development of a decision-support interface for government operators.

During critical events like natural disasters or elections, this tool would ingest a target SNS post and render the framework's output as a structured, interactive graph of evidence. This would transform the complex, multi-dimensional output of our analysis into an intuitive "auditor's report," empowering a human decision-maker to rapidly assess the credibility of a claim and take appropriate action, such as issuing public clarifications or deploying relief activities based on verified information (Shneiderman, 2020).

Furthermore, a truly effective defense requires moving beyond a standalone tool to an integrated ecosystem. Future development will focus on creating APIs to allow for partnerships with news media and SNS operators. By providing these key players with access to the framework's truthfulness judgments, we can enable them to promptly correct or contextualize malicious disinformation, thereby suppressing its viral spread and mitigating social confusion.

However, implementing this vision requires a deliberate and cautious navigation of the complex ethical and legal landscape. This will be the primary focus of our future research and development on two critical constraints:

1. **Privacy and Data Protection:** The automated collection of internet data, even for public posts, operates in a legally sensitive space. Future work must establish a rigorous compliance framework, adhering to data protection regulations like the EU's GDPR and defining a clear legal scope of application, potentially through contracts with platform operators (European Union, 2016).

2. **Copyright and Fair Use:** The use of third-party internet content as evidence for analysis places this framework at the center of the ongoing global debate around AI and copyright. Future implementation must carefully navigate the doctrine of "fair use" to ensure that the service can operate with legal and ethical integrity (Lemley, & Casey, 2021).

Addressing these constraints is a non-negotiable prerequisite for providing a service that can be used with confidence to rebuild trust in our shared information ecosystem.

Conclusion

This paper argues that the global disinformation crisis cannot be solved by creating a faster, black box "truth engine," a flawed paradigm that deepens public distrust. Instead, we present the architectural blueprint for a transparent AI auditor: **a Pluralistic Framework**.

The future of this work requires a collaborative, open ecosystem uniting academia and industry to build upon this foundation. This will enable real-world applications, such as crisis decision-support tools for governments and trust-as-a-service APIs for media organizations.

Our ultimate objective is to establish a new form of digital public infrastructure: a trusted, verifiable layer for our information ecosystem. This foundational work offers a path to protect society from the virus of disinformation and rebuild a shared reality, a challenge that is not merely technical but a societal imperative.

References

World Economic Forum, 2025. The Global Risks Report 2025. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf.

Bobby Allyn, 2022. Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn.

European Commission, 2022. Digital Services Act. Official Journal of the European Union.

Hajj, Toumi, & Zeadally, 2023. A survey of deepfake detection techniques. IEEE Security & Privacy, 21(2), 58-67.

Graves, & Adair, 2021. The State of the Fact-Checkers. Duke Reporters' Lab, Duke University.

Vosoughi, Roy, & Aral, 2018. The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Fazio, Brashier, Payne, & Marsh, 2015. Knowledge does not protect against illusory truth. Journal of Experimental Psychology: General, 144(5), 993–1002.

Chern, Aleman, Weng, & Zhang, 2023. FacTool: A tool for detecting factual errors of text generated by large language models. arXiv preprint arXiv:2307.13528.

The White House, 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Coalition for Content Provenance and Authenticity (C2PA), 2023. C2PA Technical Specification Version 1.3.

Yager, & Liu, 2008. Classic works of the Dempster-Shafer theory of belief functions. Springer.

DiResta, & Shapiro, 2024. The Weaponization of Search. The Atlantic.

Chern, Aleman, Weng, & Zhang, 2023. FacTool: A tool for detecting factual errors of text generated by large language models. arXiv preprint arXiv:2307.13528.

Grobe, & Kligler-Vilenchik, 2023. Algorithmic literacy in an age of generative AI. Journal of Communication, 73(5), 391-401.

Carlini, & Farid, 2022. The threat of adversarial attacks on deepfake detection. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (pp. 313-326).

Zimmermann, 1995. The Official PGP User's Guide. MIT Press.

Resnick, & Varian**,** 1997. Recommender systems. *Communications of the ACM*, 40(3), 56-58.

Abdar, Pourpanah, Hussain, Rezazadegan, Liu, Ghasemian, & Acharya, 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243-297.

Hacene, Huchard, & Huchard, 2022. A survey of Machine Learning Operations (MLOps). arXiv preprint arXiv:2209.09115.

Shneiderman, 2020. *Human-centered AI: Reliable, safe, and trustworthy*. Oxford University Press.

European Union**,** 2016. *General Data Protection Regulation (GDPR)*. Official Journal of the European Union, L 119/1.

Lemley, & Casey**,** 2021. Fair learning. Texas Law Review, 99, 743.