Beyond Detection: Deep Fake Verification, Editorial Accountability and Trustworthy Journalism in The Generative AI Era

Sakshi Singh[1], Dr. Ashwani Kumar[2]
[1]*Research Scholar, Department of Journalism and Mass Communication, Monad University, Hapur, Uttar Pradesh, India*
[2]*Associate Professor & Head, Department of Journalism and Mass Communication, Monad University, Hapur, Uttar Pradesh, India*

***Abstract*—Generative artificial intelligence makes it difficult to use audio visual content as journalistic evidence. However, identifying an item as synthetic does not determine if a news story is true. This conceptual article proposes an Editorial Verification and Accountability Framework (EVAF). The EVAF relates to preservation, provenance assessment, independent corroboration, risk sensitive escalation, publication decisions and post-publication corrections. Using research from journalism studies, information disorder research, media forensics and digital provenance scholarship, the article differentiates between artefact authenticity from source provenance, factual accuracy, and contextual truth. An operational account of how editors can document evidence, reveal remaining uncertainty, and assign decision responsibility was provided, including what editors should do when there are differing views among technical tools, or original files are unavailable. The framework is discussed in relation to multilingual circulation and resource constraints in India and illustrated through a hypothetical newsroom scenario. Six propositions identify potential areas for further testing. EVAF is proposed as a tentative decision support model rather than as a validated detection system or as a declaration that humans cannot make errors.**

***Index Terms*—deepfakes; journalism; verification; editorial accountability; synthetic media; provenance; multilingual misinformation**

I. INTRODUCTION

Public-interest reporting often relies upon the use of photographs, audio recordings, and video footage. Generative artificial intelligence (AI) is making the creation of fabricated audiovisual content easier to do than ever before. Furthermore, there is now an additional risk that true evidence will be dismissed as fake based on allegations of it being AI generated. These dangers could result in either false acceptance of fake content or unjustified rejection of real content, the latter is frequently referred to as the liar's dividend (Chesney & Citron, 2019; Gregory, 2022).

A detector can create a confidence level (or classification) for each item; however, it cannot determine who created the recording, the date of the event being recorded, or interpret the meaning of a translated sentence. Conversely, a real file may support an incorrect news story if its time, location, caption or sequence is incorrect. Therefore, verification involves three linked judgements: assessment of the media artefact, evaluation of the associated claim, and the editorial decision concerning publication (Thomson et al., 2022; Verdoliva, 2020).

Research has focused on the issue of uncertainty and trust of audiences (Vaccari & Chadwick, 2020), the normative limits of synthetic media within journalism (Raemy et al., 2025), changing skill sets and practices for journalists (Bendahan Bitton et al., 2026) and the real world conditions under which fact checking occurs in India (Roy et al., 2025). Furthermore, a layered visual trust model incorporates provenance, institutional authority and audience understanding (Ruiz San Miguel et al., 2026). Given these developments, it would be incorrect to state that multilayered verification is a novel concept. Rather, the specific area being studied here relates to how newsrooms use various forms of evidence to create an

accountable, appropriate and reversible editorial decision.

Deepfake research also suggests some potentially positive applications of synthesis in terms of accessibility, entertainment and the way news is presented; as long as the use of synthetic reconstruction is clearly identified and does not appear to have been created as a documentary record (Lundberg & Mozelius, 2025; Raemy et al., 2025). Therefore, the critical editorial line is not a complete ban on synthetic media but rather the relationship between representation, evidentiary value and audience comprehension.

The article asks:

1. Why is deepfake detection insufficient as an editorial verdict
2. How can the evidentiary and accountability steps of verification be specified as a usable workflow
3. How might uncertainty, harm and multilingual context change the decision to publish?

The Editorial Verification and Accountability Framework (EVAF), as proposed, will combine six established areas of activity into a formal decision record; that is, what claims have been evaluated, how each claim is supported or contradicted, who authorized the evaluation, and what conditions would require a revision of the evaluation.

II. CONCEPTUAL AND THEORETICAL FOUNDATIONS

2.1 Synthetic media, information disorder and the unit of analysis

The term deepfake does not encompass all types of misleading media. There are numerous ways to create misinformation, such as, creating a person's speech, altering someone's face, using a small part of an actual interview, or combining older footage with a new headline. Scholars define fake news and advise against combining different forms of misinformation (fabrication, satire, and manipulation) into a single category without distinguishing among them (Tandoc et al., 2018).

The distinction between misinformation, disinformation and malinformation is particularly relevant to editorial attribution. Misinformation denotes false information shared without a demonstrated intention to deceive; disinformation involves knowingly deceptive communication; and malinformation involves genuine information used in ways that can cause harm, for example through harmful disclosure or stripped context (Wardle & Derakhshan, 2017). Because intent is frequently inaccessible to a newsroom, EVAF asks editors to document observable evidence of circulation and representation before assigning an intentional deception label. The same synthetic clip might be legitimate satire in one setting and a deceptive impersonation when reposted without disclosure.

For EVAF, the basic unit of analysis is the media linked claim: a single statement made about or in conjunction with an artifact. Therefore, a single video could contain multiple claims (e.g., this person is in the video, that speech was given yesterday, the translation of the caption reflects the actual words used).

The claim-based method is also beneficial in preventing a category error when evaluating whether a file was altered versus whether a person intentionally deceived an audience. While a reporter can prove that subtitles were added to a video, they cannot always identify who added them or why. Rather than assuming the source or intent behind a verified alteration, a valid finding should document the alteration and the impact it had on the associated claim. Table 1 provides examples of how to convert these analytical differences into specific inquiries, but does not suggest that any one of the listed checks will be definitive.

Table 1. Categories of audiovisual claims, principal verification questions and supporting literature.

Media or claim category	Core verification question	Illustrative checks	References
AI-generated or face-swapped video	Did the represented event occur as shown?	Original file; independent footage; production disclosure; forensic analysis	(Raemy et al., 2025; Verdoliva, 2020)

Cloned or synthetically edited audio	Did the identified speaker say the alleged words?	Source confirmation; full recording; acoustic review; independent transcript	(Bendahan Bitton et al., 2026; Verdoliva, 2020)
Altered authentic image/video	Which substantive features or sequences changed?	Source version; frame comparison; edit history; contextual witnesses	(Coalition for Content & Authenticity, 2025; Thomson et al., 2022)
Authentic but miscaptioned footage	Are time, place and claimed event correctly attributed?	Archive search; geolocation; chronology; corroborating records	(Gregory, 2022; Thomson et al., 2022)
Re-voiced or mistranslated material	Does the translated meaning preserve the original claim?	Original-language check; two-way review; local contextual sources	(Roy et al., 2025; Shahi & Hale, 2025)
Satirical or disclosed synthetic material	Could presentation misleadingly imply documentary evidence?	Creator intent; surrounding caption; clear audience-facing label	(Raemy et al., 2025; Tandoc et al., 2018)

2.2 Gatekeeping, sociotechnical work and accountability

The process of classical fact-checking includes three components: selection of relevant evidence, comparison of that evidence to determine its accuracy, and making a public, justifiable decision about whether the claim is true or false. Automated news production redistributes tasks through algorithmic processes; however, this does not eliminate the human institutional responsibility (Diakopoulos, 2019). Studies of how deepfakes are used in news organizations also emphasize the importance of staffing, competence, organizational policies and procedures, as opposed to a simple technological fix to prevent false information from being disseminated (Bendahan Bitton et al., 2026; Weikmann & Lecheler, 2024). EVAF views verification as an organizationally coordinated effort where technology can assist in the decision-making process, but do not replace editorial authority.

Raemy et al. examine if and when journalists' application of synthetic media aligns with their professional obligations (Raemy et al., 2025). Ruiz San Miguel et al. outline the multiple levels of visual trust. EVAF answers a more specific operational question: How does an editor evaluate an evidential claim during review, record residual uncertainty, and they decide whether to publish, qualify, postpone, or reject the claim? (Ruiz San Miguel et al., 2026). This model's value stems from the decision-making structure and the accountability trail it provides, not from creating a new ethical framework or universal measure for evaluating claims.

2.3 Authenticity, provenance and contextual truth

Artifact authenticity is about whether the object being described has the production history or integrity attributed to it. Provenance refers to how you can find out more about the history of the object. Factual accuracy is about whether what is stated as fact actually happened. Contextual truth is about whether the timing, information framing, order in which it is presented, and translation allow for the same meaning to remain intact. An accurate description may contain an incorrect caption, and a lack of credentials does not mean that the item is counterfeit (Coalition for Content & Authenticity, 2025; Gregory, 2022).

The Coalition for Content Provenance and Authenticity (C2PA, 2025) provides a systematic method for signing provenance data; however, it explicitly states that provenance by itself cannot determine if content is accurate (Coalition for Content & Authenticity, 2025). This restriction is crucial to EVAF; every digital signal must be assessed based upon the particular claim under consideration and on the basis of independent evidence.

The newsroom must also differentiate between unavailable original from a suspect original. File level information can be deleted by message applications, image editing programs, and when transferring files to archives; screen shots and re-coding may hide previous processing. An intact signature typically

verifies only that the file came from the source stated, but does not provide proof that there was someone present at the time the photo was taken. EVAF will preserve the outcome of the provenance verification along with the extent of verification, i.e., which file was verified, what was confirmed, what could not be determined, and what factual statements remain unverified.

III. METHOD AND SCOPE

This paper represents a theoretical narrative review of literature. The study selected relevant peer reviewed literature related to journalism, deep fakes, visual verification, audience effects of information disorders and India's fact checking ecosystem. It has also reviewed relevant institutional and technical guidance. Sources were selected purposively for conceptual relevance, including research that establishes limits or potential alternatives to the proposed model.

This review of the literature used in this conceptual analysis was not obtained from an organized database search. There are no complete search logs, no search strings, no date-by-date screening registers, nor are there any counts of excluded articles. MacInnis identifies two forms of conceptual advancement, differentiation and integration (MacInnis, 2011). Similarly, Jaakkola suggests that researchers make the underlying logic of their model-oriented research explicitly clear (Jaakkola, 2020). In this instance, differentiation would separate the media file, the claim, and the editorial decision; whereas integration would allow for the inspection of evidence, risks, and corrections through a single record. Therefore, the novelty claim made here is intentionally limited and dependent upon additional research (i.e., comparative studies) to evaluate if the EVAF configuration would provide improved performance compared to current practices.

The analytical process is structured around four stages:

1. Identifying the media-linked claim as the focus of analysis,
2. Distinguishing between authenticity, provenance, accuracy of facts and context,
3. Organizing the newsroom obligations into six consecutive and adjustable decision steps,
4. Developing propositions for future empirical testing.

Although EVAF draws on existing scholarship, its proposed contribution lies in integrating claim-level evidence, explicit audit fields, escalation triggers and accountable editorial decisions within a single operational framework.

The study focuses on how a news organization (or fact-checking organization) processes audiovisual materials that could be potentially misleading, which are submitted by members of the general public, platforms, or directly to the organization. Both pre-publication verification and post-publication corrections fall under the purview of this research. There are no claims made regarding the prevalence of deepfakes, the performance of deepfake detection systems for Indian languages, the extent to which news organizations have adopted the use of EVAF, nor the success of EVAF's recommendations.

IV. WHY DETECTION ALONE IS INSUFFICIENT

4.1 A classifier is not an editorial verdict

Media forensics methods attempt to detect evidence of how a piece of content came into existence, as well as whether there were any changes made to it. In contrast, journalism typically requires a determination of truthfulness at the claim level. The performance of a given tool will likely vary depending upon whether or not the input media has been compressed, resized, transcended or produced through a method not well represented in the evaluation data set. Therefore, researchers caution against making the assumption that a tool's score represents a consistent value (Verdoliva, 2020). All output from a tool should include its origin, constraints and the specific question it was designed to answer.

False negatives can be used to promote false stories by giving them a platform in the media. False positives can be used as a reason to ignore actual evidence that would be of interest to the public, thus giving the liar's dividend more power (Chesney & Citron, 2019; Gregory, 2022).

Editorial risk depends upon who the allegations involve, who is affected and if the publication will result in irreversible damage. Therefore, EVAF must evaluate both erroneous acceptance and erroneous rejection simultaneously, regardless of if the same classifier is being used.

4.2 Independent corroboration must test the proposition

A more effective subsequent action would be to find the earliest accessible version, request the source document, refer to contemporary records or compare identified individuals/witnesses/locations and dates/times (Thomson et al., 2022). A repost of a single clip does not constitute independent confirmation. High quality verification establishes the connection between each check and the claim being supported, as opposed to simply using the total number of checks as an indicator of the quality of the verification.

A reverse image search or a video key frame search can find previously uploaded images or videos as well as prior descriptions of them. Geolocation can use the visual evidence of landmarks in an image to see if there is any other documentation of that place. Chronolocation could use photographs, calendars, weather reports, and/or lighting conditions at the time an image was taken. Examining the metadata and comparing each individual frame in a video will help determine whether modifications were made to the image or video. Finally, reviewing the audio will allow for testing for consistency, background noises, or who is speaking in the video. When examining translated content, the original language audio should be compared to the subtitles or dubbing by someone who is knowledgeable about the original language and its register. Searches can miss private distributions, location cues can be unclear, metadata can be destroyed or altered, and voice analysis can yield inconclusive results (Thomson et al., 2022; Verdoliva, 2020). Documenting the limitations of these methods helps prevent checklists from being perceived as proof.

An explicit status of unverified is preferred over an overly hasty conclusion regarding the verification process when there is no way to confirm the source. If verification cannot be confirmed, however, and the public interest is high enough, then a clear statement regarding the uncertainty may be warranted as long as it does not imply that the unconfirmed item is an accepted fact.

V. THE EDITORIAL VERIFICATION AND ACCOUNTABILITY FRAMEWORK

The six stages of EVAF represent a step-by-step process of decision-making support. This support enables the evaluation of how content moves from being received through to its accountable publication. While these steps may occur sequentially for record keeping purposes, they can be utilized as part of an ongoing, iterative process. If additional information becomes available, this information may lead to a re-evaluation of the origin of the content, escalation of an issue, or even a revision of a previous editorial decision. As shown in Figure 1, there are six stages within the EVAF framework, which include a log of evidence. At present, EVAF is a proposed framework rather than a validated standard; thus, it cannot replace existing professional editorial codes or legal reviews.

Evidence can help to make known the fact that there is a controversy even though it does not provide enough support for the claim made against someone. On the other hand, if we have a technically convincing finding, we may need to investigate further instead of making a public accusation. Therefore, EVAF provides two different levels of evidentiary status: an internal level and an external level. The internal level will tell us what we can currently prove; the external level will add questions regarding public interest, possible harm caused by our findings, protecting our sources, when to release our findings, and how to appropriately qualify our editorial statements.

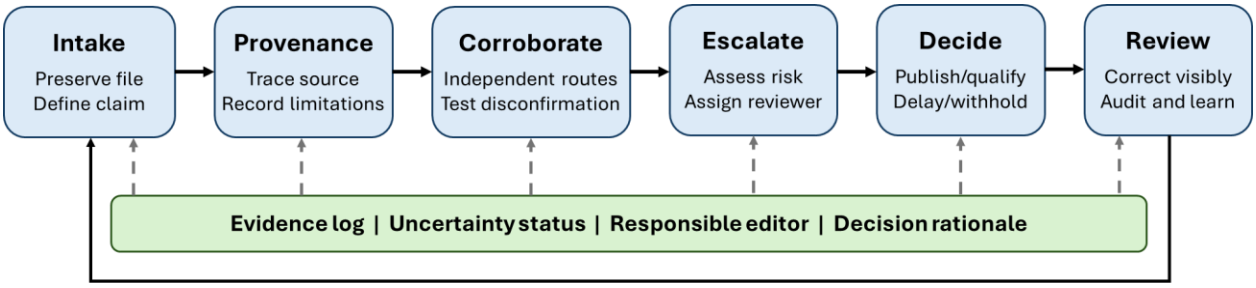


Figure 1. Proposed six-stage Editorial Verification and Accountability Framework (EVAF), illustrating claim verification, shared evidence and accountability records, and iterative review following new evidence or corrections.

5.1 Stages and evidentiary outputs

Stage 1: Receive, Preserve and Define Claims. Preserve the highest quality available version of the file, so that it does not need to be altered multiple times before being used as evidence. Record the source URL, the date of receipt, any captions, platforms, languages and other known transformations. Identify the specific factual claim being evaluated.

Stage 2: Assess the source and provenance of the Article. Find the oldest available version of the article, request the original where feasible, look at the article's metadata or credentials, and record any missing parts in the article's chain of custody. If you are unable to find the source of the article, record this uncertainty as opposed to immediately assuming that the article has been synthetically created.

Stage 3: Independent Verification. Find additional evidence to support the claims, such as, current records, translations of foreign language transcripts, alternative recordings, location data, time data, relevant documents, and credible sources. Be sure to note all conflicting results. It is important to remember the difference between verifying secondary details (e.g., the location of a street) and validating the main claim.

Stage 4: Risk sensitive escalation. Who can suffer harm as a result of accepting a false claim, rejecting a valid claim, or repeating something that isn't required? Senior review is necessary for possible election interference, violence, allegations of crimes, threats to personal safety and serious damage to reputation. When required, senior review should also include consultation with legal counsel or technical experts. Escalation cannot make up for lack of evidence; however, it does determine the amount of scrutiny and authorization required.

Stage 5: Editorial decision and explanation. Decision options include publishing with corroboration, publishing with explicit limitations, delaying for further checks, publishing a fact check of the circulating claim, or withholding the disputed content. An explanation is required that will outline what claim was tested, how the claim was tested, and what remains uncertain; AI generated, edited, out of context and unverified cannot be considered as equivalent terms.

Stage 6: Correction and organisational learning. The preservation of future evidence, correcting the article, headline, caption, and related social media posts, as appropriate, and showing the substance of the correction, are all part of correcting the issue and learning from the experience. A proportionate review of sourcing, technical assessment, escalation, and approval will be conducted. Since deepfake claims can evolve quickly, the decision to create a correction trigger is made early on in the decision-making process instead of later.

A corrective action will be taken based upon identifying what additional facts would make a material difference in the original published statement, such as an original audio/video recording, an authenticated transcript, etc. When that condition arises, then a responsible editor must review the published article and any other related content produced on the platform (clips, captions, etc.) that could still exist as part of the public record after a correction has been made to the website. It implies that when there are significant, well-documented inaccuracies, those must be corrected consistently across all of the various forms of the content controlled by the newsroom, and the correction must include documentation of why the change was made.

Table 2. EVAF stages, decision records and illustrative literature underpinning the proposed model.

Stage	Required action / checkpoint	Record, role and decision output	References
Intake	Preserve file; distinguish media item from specific claim; record language and platform.	Intake log; assigning reporter; claim statement and preservation path.	(Graves, 2016; Thomson et al., 2022)
Provenance	Trace source; check available credentials and file history; report gaps.	Provenance note; verification editor; established / incomplete / unknown.	(Coalition for Content & Authenticity, 2025; Gregory, 2022)

Corroboration	Seek independent, disconfirming and context-specific evidence.	Evidence matrix; verification lead; substantiated / disputed / unresolved.	(Thomson et al., 2022; Weikmann & Lecheler, 2024)
Escalation	Assess harm, reversibility and public interest; refer high-risk cases.	Risk rationale; senior or specialist reviewer; escalation authorisation.	(Raemy et al., 2025; UNESCO, 2023)
Decision	Choose publication status, caveats, source protection and disclosure.	Signed decision note; editor; publish / qualify / delay / withhold.	(Diakopoulos, 2019; Graves, 2016)
Review	Publish corrections, update derivatives, audit failures and training needs.	Correction log; designated editor; reopened / corrected / closed.	(Bendahan Bitton et al., 2026; Roy et al., 2025)

5.2 The evidence-and-accountability record

Across the six stages, EVAF proposes a lightweight record with ten fields: the claim, the source/file identifier, the evidentiary review, tool output and limitations, supporting evidence, disconfirming evidence, unresolved issues/uncertainty, potential adverse impacts and escalation pathways, responsible editor/publication rationale, and triggers for corrections. The form will not reveal confidential sources or release sensitive personal information.

The public facing article will only show the specific reason why the claim was verified and will not display any private data, privileged information from lawyers, or other information that could potentially harm sources. This separation between the internal documentation and a publicly accessible explanation provides transparency about how audits are conducted, without making the assumption that all forms of transparency are inherently safe. The model does not establish a single universal number threshold as a requirement for publication. Instead, the editor needs to provide documentation explaining why the evidence supporting the claim being made is appropriate for the given claim and risk associated with it, as well as who approved the article for publication.

VI. EDITORIAL DECISION MAKING AND ORGANISATIONAL ACCOUNTABILITY

Accountability is contingent upon having a designated decision maker. Reporters document the source and statement of claims, verification specialists assess technical and contextual support for claims, desk editors determine whether to publish, and publishers establish guidelines regarding correction processes and appeals. AI's role in journalism suggests that organizations will continue to maintain control over journalism practices even if they delegate some of their responsibilities to automated systems (Bendahan Bitton et al., 2026; Diakopoulos, 2019).

Synthetic can be established, even though the user does not have the intention to deceive. Authentic can be supported, even though the headline is not. Unverified can be the best conclusion for the story, even though there was a lot of interest in the story. Figure 2 shows some examples of branching logic. If the story being disputed is something that has generated public interest, then an outlet can report on the disputed story and the fact that it is unresolved, or an outlet can create a fact-checked report based upon evidence, but the outlet cannot report the dispute as if it were a verified occurrence. The decision will continue to be qualitative, dependent upon the level of risk and editorial approval, rather than being determined by a statistical classifier.

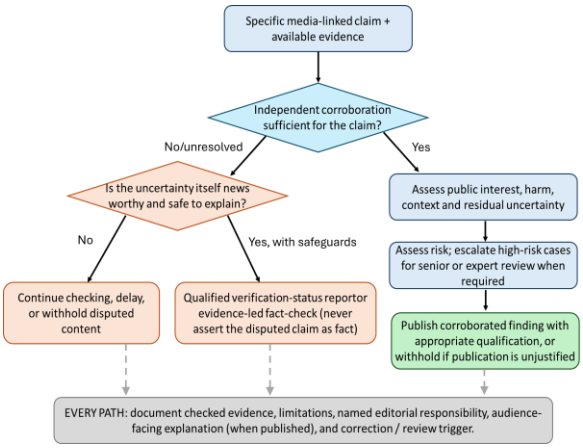


Figure 2. Revised, illustrative editorial decision pathway. An unresolved claim may warrant continued investigation, non-publication, or a carefully qualified report on the verification status itself; no branch permits an unsubstantiated allegation to be stated as fact.

An item can't be called human reviewed if there weren't appropriate checks. Public descriptions should typically avoid identifying information about sources. Responsible disclosure includes both the outcome and limitations of that outcome while protecting the individuals who provided evidence (Gregory, 2022; UNESCO, 2023).

Editorial independence is a further component of accountability. Platforms and public authorities can develop systems that allow for testing claims about where information came from, reporting false claims quickly, and accessing reliable sources of original information. A newsroom must have the ability to compare official denials of claims, platform generated warnings, and commercial detection outputs to independent sources of information. UNESCO's multi-stakeholder model supports this type of testing/verification of information, as editors will continue to be responsible for the accuracy of what they publish and must create ways for people to dispute or correct inaccurate decisions, while maintaining ultimate responsibility for those decisions (UNESCO, 2023).

VII. THE INDIAN MULTILINGUAL AND PLATFORM CONTEXT

India represents an excellent example of where to look at EVAF since there are many ways that the exact same audio-visual material can be found across various news organizations, on public social media sites, and even on private messaging services in multiple languages. For instance, Shahi and Hale study claims that were submitted via WhatsApp fact checking tip lines during the election process for several Indian states (Shahi & Hale, 2025). Roy et al. demonstrates how Indian fact checkers responded to the spread of false information regarding the elections and how they dealt with the challenges posed by language and organization (Roy et al., 2025).

A translation can change the tense of an action, leave it unclear who is speaking, or give more weight to a statement than what was originally said. Therefore, EVAF views the source language version, the translation process, and the resulting translated claim as distinct pieces of evidence. When possible, the newsroom should document who read the translation, if they took into account regional idioms, and how many times a given claim was included in the different versions of the video that were sent out. If there are multiple captions associated with a single video clip, each caption will need to be evaluated on its own merits as opposed to being labeled fake collectively.

Resource constraints are an additional factor. For example, a small local newspaper may not be able to afford the costs associated with paid forensic software or a full-time verification specialist. On the other hand, the more sophisticated processes involved in EVAF, i.e., verifying the source of the claim, obtaining at least one verified independent corroboration, assigning decision-making responsibility, and preserving files can still be used by newsrooms with varying levels of resource availability. It should be noted that the above is a theoretical proposal and not empirical evidence of the adoption of EVAF within Indian news media.

If we publish a fact check in English, it will be difficult for us to correct what is being claimed by someone who has posted a video that is in Hindi or Telugu. In future evaluations, we need to separate out the confidence people have in the source of the information from their understanding of the corrected claim. We also need to assess if translating disclosures about claims being false or unverified can provide the same level of confidence.

Future research should not assume that linguistic variation alone explains differences in interpretation. The size of the newsroom, how media literate the audience is, the geographic region, access to native language experts, whether news is distributed through urban/rural channels, and the way a platform used for messaging or public information is constructed may all impact how corrections are verified and received. Therefore, a comparative research design could provide a way to stratify outlets based on resource availability and compare groups of users. However, there is no implication that any group of users defined by language is inherently vulnerable or unreliable.

VIII. ILLUSTRATIVE APPLICATION: A HYPOTHETICAL MULTILINGUAL VIDEO CLAIM

consider a hypothetical video that was distributed with two different captions. One caption was written in English, which stated that a government representative announced a new policy yesterday. The second caption was written in Hindi and contained a more negative/inflammatory interpretation. It is important

to note that this is a hypothetical scenario and does not reflect a real-world event, an example of a field-based study, or a demonstration of the effectiveness of the EVAF system.

At intake, the reporter will archive every post made and will define two questions regarding the allegation: did the official state the alleged announcement? Does the Hindi translation accurately reflect what was said? Provenance assessment identifies the most recent available post, but it does not provide access to the initial recording. The detector produces an inconclusive outcome; due to its inability to confirm date, translation, or event context, the result is recorded as a limited piece of evidence for the claim made.

The verification team will check original transcripts, contemporaneous news reports, ask for another copy of the recording, and have a non-partisan bilingual individual review the speech to determine if there are differences between the speech as delivered, and what was transcribed. If the event actually happened, but took place months ago, and the misleading translation has added words that were never said, the editorial finding will show that while the recording of the event was accurate, the timing of the report was misleading and the translation was inaccurate. Since the story deals with potentially damaging public accusations, a senior editor will review the proposed report prior to its being published.

The resulting article will most likely state that the video is a true representation of an event, however, the information about the date and language of the video was misrepresented. In addition to this, it should be disclosed that the original video file cannot be found. The article does not need to identify the video as a deepfake. This example demonstrates the type of information that EVAF wishes to maintain, rather than demonstrating how EVAF can provide a faster service, more accurate service, or greater public trust.

IX. ETHICS, HARM AND THE RISK OF OVERCORRECTION

The process of verifying an inaccurate video can actually make things worse for the person involved, particularly if they are a member of a vulnerable group, or have been accused of something serious. News organizations must consider whether to provide their audience with a descriptive article about the issue, a single photo from the video, or some other form of context rather than simply providing them with the original video. They also need to avoid assuming that the person who initially uploaded the video was responsible for creating it, or had bad intentions. News organizations face two ethical dilemmas when deciding how to report on these issues, what is proportional in terms of what we publish and what we do not (Chesney & Citron, 2019; Gregory, 2022).

False positive allegations may restrict an audience from seeing true evidence in much the same way as false negative allegations can provide deceptive content with a credible voice. Audience research indicates that synthetic political media can affect levels of certainty and credibility. Experimental research indicates that journalistic fact checking may alleviate some negative effects of synthetic political media in certain circumstances (Dan, 2026; Vaccari & Chadwick, 2020).

These results do not demonstrate the efficacy of EVAF; rather, they emphasize the need to investigate both types of misclassifications, as well as the effects of subsequent editorial communication.

An accountable newsroom must be capable of explaining why corrections were made, without equating accountability with immunity from making mistakes. AI checked and human reviewed processes describe how things are done, but they do not provide proof that a published article is accurate. When a conclusion cannot be reasonably drawn, well-defined uncertainty can often offer greater insight than a definitive, overconfident binary classification.

X. PROPOSITIONS AND FUTURE RESEARCH AGENDA

Six propositions can guide empirical assessment.

P1: workflows integrating source provenance with independent corroboration will yield better claim-level classification and justification than detection only workflows.

P2: disclosure of specific evidence and uncertainty will increase audience comprehension and confidence calibration, although immediate publisher trust may not rise.

P3: clearly identified editorial responsibility will increase perceived accountability compared with vague attribution to an algorithm.

P4: the trust consequences of false-positive and false-negative editorial errors will differ by claim type and public-interest stakes.

P5: disclosure wording will interact with language and platform context to influence understanding.

P6: risk-sensitive escalation will reduce serious avoidable publication mistakes but may increase verification time.

Table 3 relates the hypotheses to measures of outcomes and possible experimental design. For instance, decision quality may be measured through correct claims against independently verified ground truth and blind evaluation of written justifications, decision quality cannot simply be defined through self-reported confidence. Studies of audience response should separately assess the degree of trust in the article, trust in the publisher, intention to share the content, and ability to identify the evidence that supports a conclusion.

An experimental study of audiences, pre-registered, could examine the predicted associations using a factorial design with three levels of manipulation: media condition (authentic vs. selected editing vs. synthetic vs. authentic-but-contextually inaccurate); disclosure condition (no labels vs. short status labels vs. specific descriptions of evidence); and explanation of validation process (present vs. absent). Stimuli used in the experiment will have been validated for establishing ground truth, and participants will be randomly assigned to one of the experimental conditions.

Potential outcome measures may include perception of authenticity, factual accuracy, source credibility, trustworthiness toward publishers, calibration of confidence, recall of evidence, and willingness to distribute information. Translation should be tested prior to copying from one language to another, and the contextual environment of the platform where the study takes place as well as participants' level of media literacy could serve as moderating variables. Sample size calculations must be completed for the factorial design, and researchers must pay attention to possible interactive effects instead of assuming that each combination can be analyzed appropriately. This type of experiment will assess audience reactions to EVAF-based explanations but does not validate the complete process of producing news.

Table 3. Testable EVAF propositions, outcome measures and research grounding.

Proposition	Expected relationship	Suggested design / outcome	References
P1: Evidence integration	Layered assessment outperforms detection-only review on defensible conclusions.	Randomised editor task; claim-level accuracy and blinded rationale quality.	(Thomson et al., 2022; Verdoliva, 2020)
P2: Uncertainty	Evidence-specific disclosure improves comprehension and calibrated confidence.	Audience experiment; explainable evidence recalls and calibration.	(Dan, 2026; Vaccari & Chadwick, 2020)
P3: Responsibility	Named editorial responsibility increases perceived accountability.	Vignette experiment; accountability ratings with trust separate.	(Graves, 2016; Raemy et al., 2025)
P4: Error asymmetry	False-positive and false-negative decisions have different reputational consequences.	Contrasting verified scenarios; publisher and claimant trust.	(Chesney & Citron, 2019; Gregory, 2022)
P5: Translation	Label effectiveness varies by language, platform and claim formulation.	Multilingual experiment; interpretation and recall accuracy.	(Roy et al., 2025; Shahi & Hale, 2025)
P6: Escalation	Risk-sensitive escalation reduces severe errors but adds decision time.	Comparative newsroom workflow; error severity and elapsed time.	(Bendahan Bitton et al., 2026; Weikmann & Lecheler, 2024)

The evaluation can include a balanced sample of independently verified authentic, synthetic and contextual examples provided to newsroom teams via existing processes or an EVAF based worksheet. Independent assessors will then rate the accuracy of decision making for claims made, the completeness of the documentation (trail) used to support those claims, the severity of any preventable errors, and record the amount of time and resources spent to produce each article. To ensure that there are no confounding factors that affect the comparison, the order of cases, the complexity level of the articles, and the experience level of the reporters must be factored into the study design.

A complementary study of the newsroom may examine the differences between what's written in policy documents versus what actually occurs in everyday practices; determine who has the authority to sign off on high-risk audio-visual claims, and find out what occurs once there has been a correction made. Recruitment, confidentiality protections, transparency in analysis, and obtaining ethical approval will be needed when using interviews and observations. Prior to utilizing literal translations for multilingual labels, these labels should be cognitively pre-tested.

Interviews would demonstrate who employees believe are responsible for making decisions. Observations of employees and de-identified decision records could show whether there is an increase in pressure to meet deadlines that leads to an increase in escalated decisions. Researchers should pay special attention to false-positives when evaluating the accuracy of media reports. False-positives occur when employees reject a legitimate recording before publishing it. These errors are more difficult to detect if the media organization does not publish a correction. An independent reference set will be necessary to estimate the number of false-positives.

XI. LIMITATIONS

EVAF is a conceptual synthesis of a variety of scholarly works and guidelines. It utilizes a purposeful literature search, not an exhaustive one. Some pertinent studies, technological developments, or individual newsroom practices may be overlooked. Since EVAF was never compared to another editorial workflow, it would be incorrect to assert that EVAF enhances accuracy, velocity, confidence, or minimizes damage. The risk pathway is based upon qualitative analysis and provides flexibility in establishing editorial thresholds. In smaller newsrooms, documenting the process can conflict with the need to report on urgent issues; furthermore, the availability of source documents, independent sources, and/or translator support may not exist.

The hypothetical case study demonstrates how we can use the distinctions made within EVAF. However, it does not provide evidence of the validity of these distinctions. It is required to test the model to see if it provides enough distinction from other models such as layered verification and visual trust models, and whether the proposed records provide a form of accountability that is useful, or just more red tape. Differences in institutions and laws around the world may necessitate that the model is adapted for each country based on those factors, beyond what has been illustrated by the Indian examples presented here.

XII. CONCLUSION

Deepfaking validation is better viewed as a judgment of proof, interpretation in context, and editing responsibility, as opposed to a battle of human judgment against an automated rating system. Determining whether an object is original (authenticity), how it was created (provenance) and whether the statements regarding the object are true (accuracy of claims), all involve similar but separate judgments. Original recordings can be misrepresented, and synthetic/edited recordings can be presented transparently and not deceive. If a recording cannot be validated, it should not be categorized as fake without sufficient evidence.

The proposed EVAF categorizes these distinctions into six dynamic stages that can be revised. In addition, the EVAF provides an evidence log, clear uncertainties, risk-based authorization and responsibility for correcting mistakes. The primary purpose of the EVAF is to provide a practical tool to help news outlets demonstrate why they made a particular decision, especially when there are multiple competing claims on one story. While the EVAF is currently considered to be provisional, its effectiveness will ultimately be determined through empirical evaluations of decision-making processes, implementation issues, audience understanding, and fair distribution among newsroom staff.

REFERENCES

[1] D. Bendahan Bitton, H. Ötting, A. Godulla, and C. P. Hoffmann, "Implications of deepfake technology for journalistic practice: A technology at the crossroads of challenges and opportunities," *Journalism*, 2026, doi: 10.1177/14648849261418328.

[2] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Review*, vol. 107, no. 6, pp. 1753–1820, 2019, doi: 10.15779/Z38RV0D15J.

[3] Coalition for Content Provenance and Authenticity, *Content Credentials: C2PA Technical Specification*, ver. 2.2, 2025.

[4] V. Dan, "Deepfakes as a democratic threat: Experimental evidence shows noxious effects that are reducible through journalistic fact checks," *The International Journal of Press/Politics*, vol. 31, no. 3, pp. 525–550, 2026, doi: 10.1177/19401612251317766.

[5] N. Diakopoulos, *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press, 2019.

[6] L. Graves, *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*. Columbia University Press, 2016.

[7] S. Gregory, "Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism," Journalism, vol. 23, no. 3, pp. 708–729, 2022, doi: 10.1177/14648849211060644.

[8] E. Jaakkola, "Designing conceptual articles: Four approaches," AMS Review, vol. 10, pp. 18–26, 2020, doi: 10.1007/s13162-020-00161-0.

[9] E. Lundberg and P. Mozelius, "The potential effects of deepfakes on news media and entertainment," AI & Society, vol. 40, pp. 2159–2170, 2025, doi: 10.1007/s00146-024-02072-1.

[10] D. J. MacInnis, "A framework for conceptual contributions in marketing," Journal of Marketing, vol. 75, no. 4, pp. 136–154, 2011, doi: 10.1509/jmkg.75.4.136.

[11] P. Raemy, D. Bendahan Bitton, H. Ötting, C. P. Hoffmann, A. Godulla, and M. Puppis, "Deepfakes and journalism: Normative considerations and implications," Journalism Studies, vol. 26, no. 14, pp. 1724–1744, 2025, doi: 10.1080/1461670X.2025.2547300.

[12] E. Roy, K. A. Laskar, and M. Reyaz, "Retroactive gatekeepers: Fact-checkers, electoral disinformation, and audience engagement in India," Emerging Media, vol. 3, no. 4, pp. 582–600, 2025, doi: 10.1177/27523543251395190.

[13] F. J. Ruiz San Miguel, A. Torres-Toukoumidis, M. Hinojosa-Becerra, and I. Marín-Gutiérrez, "From daguerreotype to deepfake: A layered model of visual trust in AI-era photojournalism," Frontiers in Communication, vol. 11, Art. no. 1959537, 2026, doi: 10.3389/fcomm.2026.1959537.

[14] G. K. Shahi and S. A. Hale, "WhatsApp tiplines and multilingual claims in the 2021 Indian assembly elections," Online Social Networks and Media, vol. 49, Art. no. 100323, 2025, doi: 10.1016/j.osnem.2025.100323.

[15] E. C. Tandoc, Jr., Z. W. Lim, and R. Ling, "Defining 'fake news': A typology of scholarly definitions," Digital Journalism, vol. 6, no. 2, pp. 137–153, 2018, doi: 10.1080/21670811.2017.1360143.

[16] T. J. Thomson, D. Angus, P. Dootson, E. Hurcombe, and A. Smith, "Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities," Journalism Practice, vol. 16, no. 5, pp. 938–962, 2022, doi: 10.1080/17512786.2020.1832139.

[17] UNESCO, Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information Through a Multistakeholder Approach. UNESCO, 2023.

[18] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," Social Media + Society, vol. 6, no. 1, pp. 1–13, 2020, doi: 10.1177/2056305120903408.

[19] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[20] C. Wardle and H. Derakhshan, Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe, 2017.

[21] T. Weikmann and S. Lecheler, "Cutting through the hype: Understanding the implications of deepfakes for the fact-checking actor-network," *Digital Journalism*, vol. 12, no. 10, pp. 1505–1522, 2024, doi: 10.1080/21670811.2023.2194665.