

An aerial photograph showing a wide, muddy river flowing through a valley. The river is filled with sediment and debris. On the left, a steep, forested hillside shows signs of erosion and landslides. In the distance, more mountains are visible under a cloudy sky. A small village is visible on the right side of the river.

WITNESS

FIVE LESSONS

from the Nepal floods, as AI worsens
audio-visual truth in climate disasters

What the response to a Himalayan flood tells us
about AI detection, and about the companies that
now supply both the fakes and the tools to spot them

TABLE OF CONTENTS

LESSON 1

Not all the AI fakes were made for Nepal 09

LESSON 2

Post-hoc detection is weakest exactly where 13
climate disasters live

LESSON 3

The increasing role of companies' detectors, and 17
their responsibilities

LESSON 4

Despite AI, the verification that mattered in Nepal 23
was human

LESSON 5

The state response to AI content is where this 26
becomes a rights issue

WHAT SOLUTIONS ARE NEEDED 30





On 26 August, a glacial collapse high on the Nepal-Tibet border sent a wall of ice, rock and water down the Bhote Koshi and Trishuli valleys. More than 1,380 people were killed in Nepal and over 5,100 remain missing, with 43 dead and more than 500 missing on the Tibetan side. Within hours, social media filled with footage of the disaster. Some of it was real. A great deal of it was not.

Five months earlier, launching the [World Disasters Report 2026](#), The International Federation of Red Cross and Red Crescent Societies (IFRC) Secretary General Jagan Chapagain said that in every crisis he had witnessed, "information is as essential as food, water and shelter", and that when it is manipulated it can deepen fear, obstruct humanitarian access and cost lives. The report frames harmful information as a humanitarian crisis in its own right. Nepal is a recent climate disaster to test that framing against an industrialised supply of synthetic media.

It is also the first test of company detectors available worldwide against emerging regulatory standards. On 2 August, three weeks before the flood, the [California AI Transparency Act](#), a state law aimed at increasing transparency around AI-generated content, and the [EU AI Act](#), the European Union's comprehensive framework for regulating AI, both began requiring generative AI providers to offer a public tool that lets anyone check whether content came from their systems. While these regulations don't bind Nepal, they can be seen as a catalyst for detection tools rolling out globally, as industry needs to adapt the ones available worldwide to comply with these emerging rules.

What makes the Nepal case distinctive is not the volume of fakes but that some of them contested the cause rather than the event. Agence France-Presse (AFP), Nepal Fact Check and VERA Files debunked AI videos attributing the flood to [a Chinese-built dam](#) and to [a missile strike on a dam](#). [TheJournal.ie](#) found such clips circulating in posts calling the disaster man-made. Synthetic media was being used to dispute a scientific account of a glacial collapse, not merely to embellish its aftermath.

Doubt ran in every direction. Among the most important authentic footage of the disaster was [security-camera video from the Gyirong border post](#), which fact-checkers had to confirm was real. [Full Fact debunked claims](#) that the whole flood was a hoax. None of this is denial of the kind documented in Iran and Gaza, where dismissing authentic documentation as synthetic has been [a deliberate strategy, but the influx of AI content speaks to the uncertainty and doubt that the increase of this type of content produces within the overall information ecosystem](#).

And the tools that settled what was real were increasingly built by the same companies whose models made the fakes. That is the through-line here. Google, OpenAI and Meta had tools capable of identifying content generated by their own systems. In the content from this disaster, the technical capability to identify synthetic content largely existed; the governance around it did not. Watermarks worked, detectors read them, labels appeared. But the capability sat with a handful of companies, worked only when someone external went looking for it (external fact-checkers, not their own provenance-based detection), and reached a government with no published criteria for how AI content could be removed.

As WITNESS put it at a [convening hosted with the Forum on Information and Democracy at IFRC headquarters](#) in July, naming harms is not the same as governing them, which is what we are increasingly seeing in this crisis and beyond: AI transparency regulations are not rising to the scale of the harms.

Lastly, it is important to note that the most consequential information failure in Nepal was not synthetic. [Mongabay reported](#) that the official SMS alert system arrived late, in Nepali only, and without actionable, location-specific guidance, while informal community warnings saved lives. The report also cites a [study following a 2021 disaster](#) that killed 120 people, which called for binding procedures linking forecasts to action. The recommendation remains unimplemented.

At WITNESS, we run the [Deepfakes Rapid Response Force \(DRRF\)](#), which since 2023 has connected journalists, fact-checkers and human rights defenders with detection experts to support them with challenging cases of potential AI content in diverse global contexts. Through this work, WITNESS has observed trends and challenges of AI detection across elections, conflict and protests settings. The Nepal cases analysed by global fact-checkers engage with the same issues and confirm some patterns, complicate others, while pointing to a few completely new developments.

A note on terms. AI detection covers two different things, and the distinction matters throughout.

Post-hoc detection, as used here, refers to detection tools that analyse a file after it has been created and circulated. These tools examine the content itself for traces that suggest it may have been generated or manipulated by AI, and produce a probability-based assessment.

Provenance-based detection, sometimes referred to as company-based classifiers, reads a signal attached at or near the point of creation: an invisible watermark, a fingerprint, or content credentials recording how a file was made and edited.

To use food as an analogy: post-hoc detection infers the ingredients and the recipe by examining the dish itself, while provenance-based detection reads the recipe the chef wrote down and sent with it. The first examines the artifact itself for evidence of how it was made; the second reads what the file records about its own making. Post-hoc detection deals with uncertainties. Provenance-based detection, attached at the point of creation, can carry more certainty, provided the signal is readable and what it means is communicated well.

Marking refers to the technique of embedding provenance signals (such as visible or invisible watermarks, digital fingerprints, or content credentials, etc.) into AI-generated media at or near the point of creation, especially within Generative AI models, the focus for this report. Marking can be applied to capture technology such as cameras or editing software that uses both AI and non-AI capabilities.

What follows is five lessons from how the Nepal cases were investigated and labelled, and what we think should change as a result.



LESSON 1

Not all the AI fakes were
made for Nepal



Some of the fake content identified in this crisis was not made originally for Nepal. AFP found one elephant rescue clip [circulating in Philippine posts in early August](#) and again [during flooding in Assam](#), weeks before the collapse. [VERA Files](#) traced an AI dam-collapse video to copies from May.

This is a standing inventory of generic disaster content, re-captioned and recontextualized for each new catastrophe. Its spread and incentives continue across disasters and crises because most social media platforms' content policies encourage engagement and monetization - including with sensationalised content (whether AI or not) throughout this disaster.

Re-detecting the same material across multiple languages and platforms wastes the scarcest resource in the response: the time and attention of local fact-checkers and investigators. Platforms already resource shared content-matching for infrastructure for priorities the industry has prioritized, such as Child Sexual Abuse Material (CSAM) and terrorist content. Applying this same capability would allow a debunked clip to carry its context whenever it resurfaces, ensuring users encounter relevant information immediately during an emergency. The goal should be a helpful label, not a takedown.



An aerial photograph of a coastal village. In the foreground, a river or canal flows through the settlement. On the left bank, a long, low building with a grey tiled roof and a blue and white patterned wall sits on stilts. The right bank is densely packed with stilt houses of various colors (white, yellow, blue) and roof types (corrugated metal, tiled). Some houses have laundry hanging outside. The background shows a lush green forested hillside under a clear sky. A large, semi-transparent circular graphic is overlaid on the right side of the image.

LESSON 2

Post-hoc detection is weakest exactly
where climate disasters live

Nearly every piece of synthetic content coming out of this crisis features water, buildings, dams, glaciers or animals. None of the tells fact-checkers relied on involved a human face. [NewsMeter](#) pointed to parked vehicles dissolving into floodwater rather than being carried away, and buildings appearing to melt. Similarly, [TheJournal.ie](#) noted trees dissolving and road signs dragged along by unseen forces, while [AFP](#) flagged an elephant whose skin wrinkled only in isolated patches.

This nuance matters because [detection has largely been built around faces and not yet prepared to deal with content related to the environmental crisis](#). Much of the research, the training data and the public imagination of a “deepfake” centres on a manipulated person's likeness. Environmental and disaster content presents a different problem. Its ground truth is physical, and its anomalies manifest as failures of physics rather than human anatomy, which is how the next generation of video models are evolving to become more sophisticated in producing realism.

These shortcomings in the design of AI detection tools significantly impact the work of the fact-checkers. [WITNESS](#) had already observed how often publicly available AI detection tools provide unreliable results, and how one simple change, such as cropping an image, can lead the detector to give an inaccurate result. In Nepal, where fact-checkers turned to commercial “post-hoc” detection tools, results were similarly uneven.

On a compilation of three clips shared as CCTV (security camera) footage, [one detection tool rated the first clip over 89% likely to be AI, the second at 54.9%, and could not reach a view on the third](#). Another outlet ran what appears to be the same compilation through a different tool and got more than 99%. Near-identical material, and confidence ranging from inconclusive to near-certain.

The failures of AI detection tools make the work of fact-checkers even more difficult, as they already verify enormous volumes of content under huge time pressure. [WITNESS' TRIED](#), Truly Innovative and Effective AI Detection Benchmark, a framework to evaluate and improve AI detection tools, presents an in-depth analysis of AI detection challenges in real-world settings.

Factors common to content shared on social media, such as low quality, high compression, background music or cross-talking, affect reliability of AI detection results.

In roughly one third of the cases supported by the DRRF, detection tools have failed to give reliable results for these reasons.



Screenshot of a Facebook video falsely shared as footage of the Nepal floods, identified by AFP.



LESSON 3

The increasing role of companies' detectors, and their responsibilities

The clearest verdicts came from watermarks and signals AI models embedded. AFP ran the most viral fake, a torrent breaching a riverbank onto pedestrians and cars, through Google's SynthID detector, which confirmed it came from Google's tools; France 24 counted one version at more than seven million views. For fabricated before-and-after images of the town of Timure, OpenAI's verification tool found Google's SynthID watermark on a piece of Google generated content, a small but real example of one company's tool reading another's mark and in truth another company's AI-generated content. Meta's online detector flagged an AI dam-collapse video and an elephant rescue clip likely made with Meta's own tools.

This is what WITNESS has long argued: durable provenance embedded at the point of creation is a far firmer foundation for verifying content and establishing or challenging its authenticity than probability from post-hoc detection technology. It worked in Nepal precisely where "post-hoc" detection was unreliable, and it worked regardless of subject matter, which makes it most valuable for the non-facial content where detection is weaker. But it also shows how much of the verdict now sits with a handful of companies, and how uneven their stewardship is.

These tools were three weeks old as legal obligations when the flood hit. On 2 August, the California AI Transparency Act and the EU AI Act's Article 50 both began requiring generative AI providers to offer a public tool that lets anyone check whether content came from their systems. WITNESS and Indicator tested what actually existed in the weeks between that deadline and the disaster. Seven of thirteen providers had no public detector at all. Those that existed mostly recognised only content produced by their own AI models, staying silent on other companies' content even when both use industry-wide credentials such as Coalition for Content Provenance and Authenticity (C2PA). Several were fooled by ordinary handling, a crop or a photograph of a screen. So the signal that settled several Nepal cases was available for a minority of the models in circulation, and silence elsewhere was not evidence of anything.

Each detector identifies mainly marks from their own Generative AI models:

Cross-reading, as with OpenAI's tool reading Google's watermark on a piece of Google generated content, is the exception rather than the design. As highlighted in a recent article on Indicator, detectors need to evolve from the minimum bar of detecting their own, unedited content, as well as becoming more intuitive and nuanced providing stronger explanations regarding the file's likely origin and edits.

Companies don't automatically detect and label their own marks:

Meta's detector is an online tool that fact-checkers consult separately from the feeds where content spreads, and as reported when it was previewed in July, it covers content from Meta's newest models. The elephant clip, likely made with Meta's tools, was shared more than 800 times on Facebook [archive here] before a fact-checker ran it through Meta's own detector. Meta only placed a misinformation label on it based on the work of AFP and NewsMeter not based on their own systems detecting their own generation but outsourcing the labour to fact-checkers.



An absent signal proves nothing, and is not a verdict:

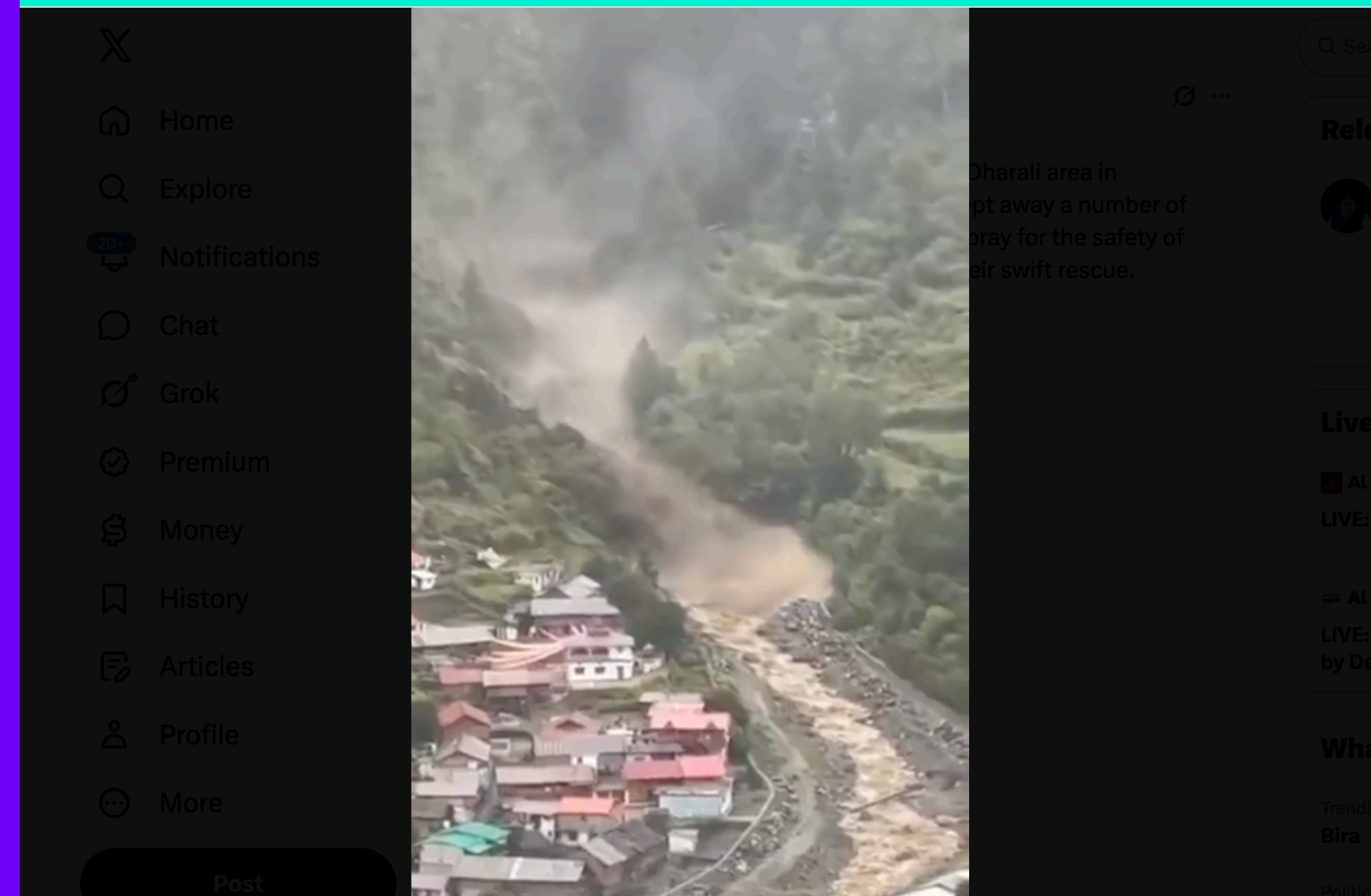
A watermark or provenance detector that returns nothing establishes only that one company's mark is absent, and many cannot read industry-wide content credentials such as C2PA. There is no prescriptive form for capturing provenance information, and no prescriptive process for validating and surfacing it, though legislation is beginning to narrow the path. The resulting variation means an absent signal says nothing in particular about a file's source or history.

Open source models can be run with no mark at the point of AI Generation, which often falls outside transparency requirements entirely. As people come to expect a clean company answer, absence gets misread as authenticity and inconclusive results get misread as proof of fakery. Lastly, this case surfaces the need for more integration and interoperability across current and incoming detectors and labels, all in order to avoid potential misreadings.

The same signal is surfaced differently across different companies:

Across the same wave of fabricated flood imagery, Australian Associated Press (AAP) recorded three different labelling vocabularies: Meta marked a before-and-after post as "AI content", YouTube flagged a video's sounds and visuals as altered or generated with AI, and an X post carried a "Made with AI" tag. The signals reached the feed. What they meant, and how much of each post they covered, was left to the reader. Establishing a standardized vocabulary across all platforms is essential so that users can immediately understand the provenance, nature, and extent of modifications in synthetic content without ambiguity.

With rights guardrails in place, for example to protect privacy, provenance should be on by default across model generators and distributors of content, readable across companies rather than only by its maker, and visible where content spreads rather than on a separate website. No single signal carries this alone: watermarks are not always resilient or recognised by others, credentials can be stripped, and marking has to be multilayered. Disclosure also needs a shared vocabulary: consistent fields, consistently placed, distinguishing what was generated from what was edited and saying what a finding rests on.



Screenshot of a video on X/Twitter that actually shows a deadly mudslide in India in 2025, identified by AAP as footage that was falsely shared as being from Nepal in 2026.



LESSON 4

Despite AI, the verification that mattered in Nepal was human

The Gyirong security-camera footage was among the most important authentic evidence of the disaster, showing the moment floodwaters destroyed the border crossing, and it was questioned as AI-generated because the colours of trees, buildings and trucks shift partway through the clip. Here, verifying the footage was not about AI detection tools but rather methods of verification and Open-Source Intelligence (OSINT) investigation.

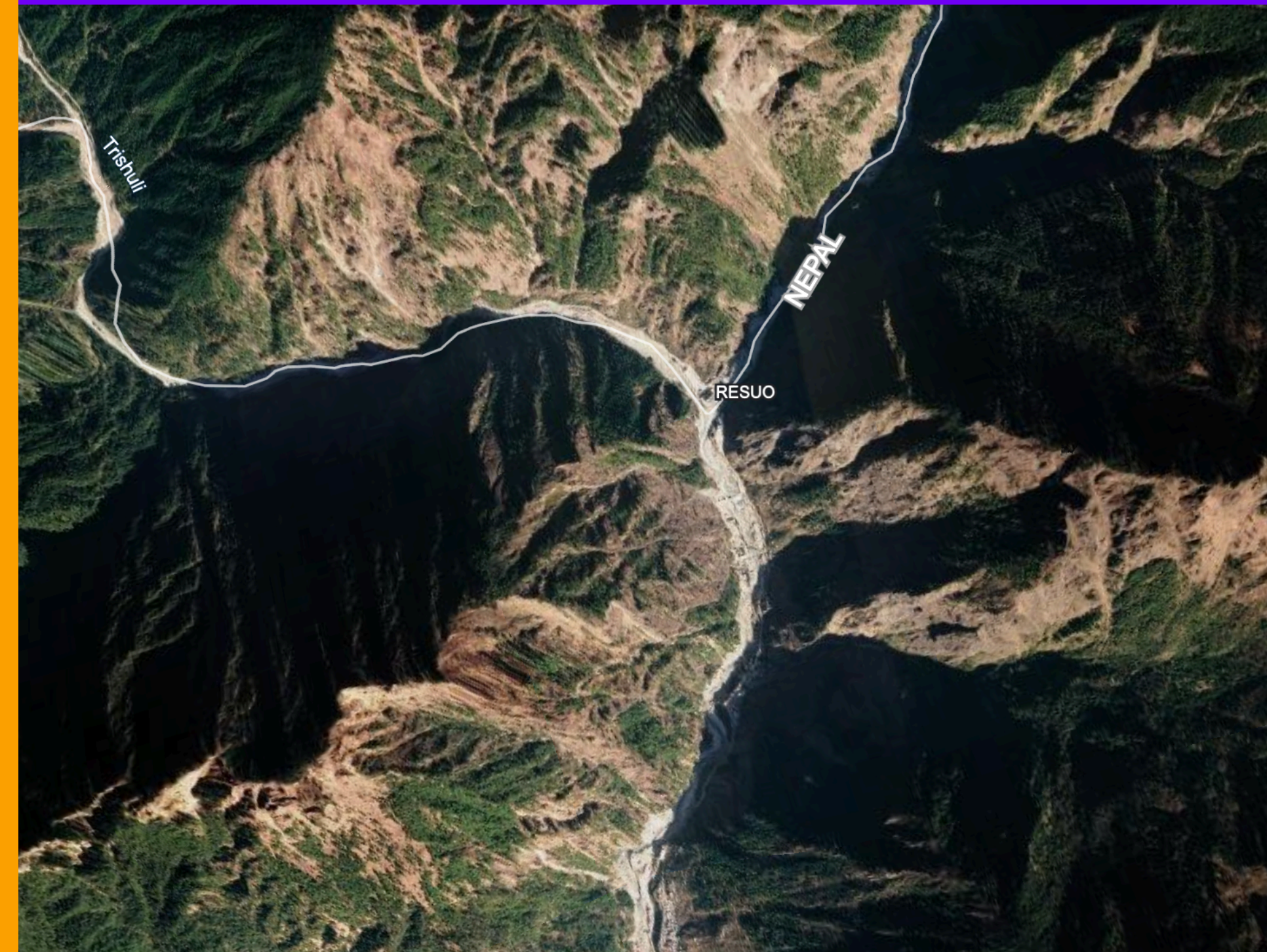
The Associated Press published the footage, and Snopes traced the circulating version to a logo from Annapurna Post, a Nepali outlet that published it on 26 August, noting that drone footage afterwards confirmed the structure had been destroyed. BBC Verify journalist Shayan Sardarizadeh, who was repeatedly asked whether the clip was AI, set out the reasoning: the camera position could be geolocated, the features matched satellite imagery of the location, and other footage from the same crossing corroborated it. His analysis concluded the colour shift appears to be an artefact of how the clip was captured from playback software.

Set beside the AI clips styled as CCTV, one carrying a timestamp that stayed fixed while the scene and camera angle changed around it, and the shape of the problem is clear. The signals of authenticity have come loose from authenticity in both directions: fabricated footage borrowing the grammar of security cameras, and real security footage looking suspect because of how it was re-recorded. Neither necessarily required AI detection methods for verification.

Similarly images of Timure did not match satellite imagery of the town. Footage shared as a glacial lake outburst flood in Rasuwa, captioned in English as "This is a GLOF", was traced by Nepal Fact Check to the Columbia Glacier in Alaska, where the pilot who filmed it had posted it days earlier. Claims about the cause were checked against the US Geological Survey's finding of a glacial collapse.

Environmental events often have physical ground truth that can be verified against satellite imagery, seismic records and river systems, that a synthetically generated visual of a person doesn't necessarily have. The work of WITNESS' DRRF emphasises the complementary relationship between AI detection and open-source investigation, which should be deployed in parallel rather than in sequence. Detection is one input to a broader verification process, and the extent of contextual information that environmental data provides is why we expect climate and disaster cases to lean on the investigative side, for both AI-generated and miscontextualised content.

It also means the capacity that mattered most in Nepal was human. Local fact-checkers and open-source investigators did work no detector could. Nepal Fact Check was debunking footage of its own country's disaster while international outlets worked the same material in parallel, and that capacity is the least funded part of the response. It should be resourced accordingly.



Google Earth visualization of one of the areas most affected by the floods in Nepal.



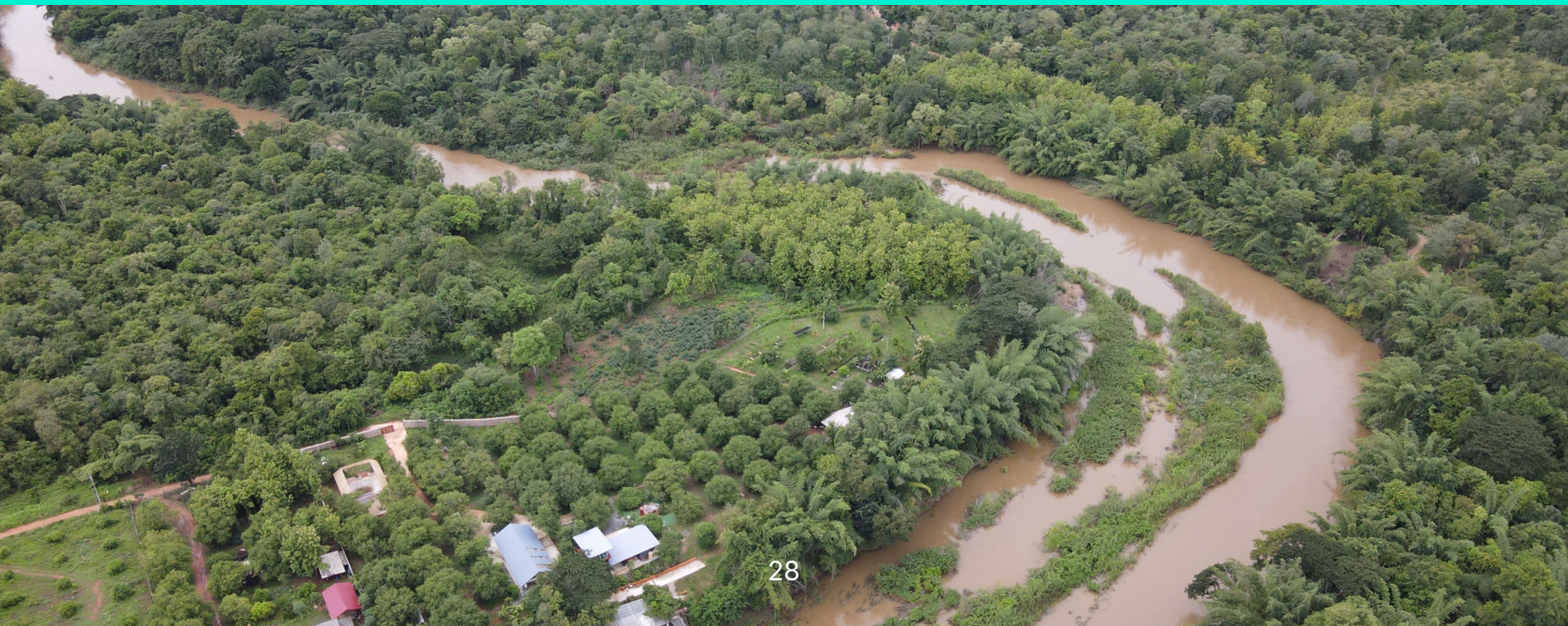
LESSON 5

The state response to AI content is where this becomes a rights issue

Immediately after the disaster, Nepal's authorities asked Meta and TikTok to remove AI-generated videos and false posts, and to reduce the reach of sensational content. At the same time, the Nepal Press Council opened a reporting portal and instructed 72 users to take material down.

Both categories rest on judgements for which no criteria have been published. "Excessively sensational" has no definition. And whether something is AI-generated is, as this piece has set out, frequently a contested reading of post-hoc detection rather than a settled fact, which means a takedown standard is being built on the least reliable part of the stack. The IFRC's own report supplies the test. Its chapter on navigating regulation, rights and societal resilience warns that vague or overly broad definitions of harmful information risk suppressing journalism, silencing dissent and curtailing civil society, and that responses must be grounded in legality, necessity and proportionality, the three-part test under Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

The Nepali state has recent reason to be careful here. A ban on 26 social media platforms in September 2025 helped trigger protests in which at least 76 people were killed and which brought down the government. The ban was withdrawn within days, and the interim administration that followed went on to withdraw the Social Media Bill 2025 altogether. A request to remove a category of content defined as "AI-generated" or "excessively sensational", made informally and without published criteria, is the kind of discretion that history argues for defining carefully.



In a crisis, a contextual label preserves documentation that removal destroys. And where states ask platforms to act, the request and its outcome should be published.



A person with dark hair, wearing a black leather jacket and shorts, is sitting on a pile of red bricks in the foreground. They are looking down at a smartphone in their hands. In the background, there is a multi-story white building that appears to be damaged or under construction, with some windows missing. The building is situated on a riverbank. The river is filled with muddy water and debris, including large piles of brown material. A forested hill is visible in the far background. The overall scene suggests a post-disaster or environmental impact situation.

**WHAT
SOLUTIONS
ARE NEEDED**

01

Matching that produces useful labels:

Debunked crisis footage should carry that context when it resurfaces, rather than being re-investigated for each new disaster.

02

Tools that show their uncertainty:

Post-hoc detection should report confidence in ways that resist being flattened into binary results, and be benchmarked publicly on non-facial, compressed, multilingual, global-majority content. Provenance tools should explain a file's likely origin and edits rather than returning a bare verdict.

03

Provenance by default, searchable across companies, visible and detectable where and when content spreads:

Examples of labels occurring on company platforms after fact-checkers used their company detectors cannot cut it. Additionally, marking has to be multilayered, and disclosure needs a shared vocabulary: consistent fields, steadily placed, distinguishing what was AI-generated from what was AI edited and saying what a finding rests on. Communicating about AI content can be delicate, but an exercise that needs investment and literacy education and research.

04

Resource community based investigators and fact-checkers:

Local fact-checkers and open-source investigators will always be crucial to building an understanding of the facts. Because context is so important - in this case they did the work no detector could - and they are the least funded part of the response.

05

Labels instead of or before content removal, and publish the requests:

In a crisis, contextual labels preserve documentation that content removal destroys, and where states ask platforms to act, the requests and their outcomes should be public.

None of this is far from what the humanitarian sector is already asking for. The [World Disasters Report](#) calls on technology platforms to prioritise authoritative information from trusted humanitarian, health and local actors in crises, to provide low-bandwidth, multilingual and locally relevant tools, and to moderate harmful content transparently. It asks states to invest in evidence-based regulation and to support local data systems, strengthening transparency and accountability.

The lesson of Nepal is not that the fakes were convincing. It is that the means of establishing what was real now sit largely with the same companies whose models made the fakes. They hold the capability. What they have not done is resource it, deploy it where content spreads, or invest in it at the moment crises demand.



CREDITS

Written by:

Mahsa Alimardani, Bruna Martins dos Santos,
Zuzanna Wojciak, Jacobo Castellanos, and
Georgia Edwards

Reviewed by:

Mahsa Alimardani, Bruna Martins dos Santos,
and Jéssica Ribeiro

Supervised by:

Nkem Agunwa

Layout by:

João Oliveira

WITNESS

witness.org