

Verify

Real-time moderation for live streaming

Verify

Real-time moderation for live streaming

Open Feed Network, Inc. · candortrustandsafety.com *Every claim maps to a dated behavioral proof in the public evidence record.*

The problem

Live video is moderation's hardest case. Harm that happens in real time cannot wait for a review that arrives hours later. Most platforms respond by either not offering live streaming or offering it unmoderated. Verify exists so a platform can offer live streaming without walking away from its safety obligations.

Design principle

The same fail-closed philosophy that governs the rest of the architecture applies here: live streaming is offered **because** the moderation covering it exists, not as a feature bolted on with moderation as an afterthought.

How it works, in principle

Verify sits on a shared real-time pipeline rather than a bespoke one-off: an active stream is captured and segmented as it happens, each segment is transcribed in near-real-time, and the transcript is fanned out to the relevant safety engines — terrorist-content detection, stream-level moderation, and live distress signals. Where a threshold is crossed, the receiving engine escalates to human review or triggers a stream-level action. The judgment of whether transcribed speech is genuinely harmful is not a separate, untested system — it reuses the same independently verified engines proven elsewhere in the evidence record.

A note on methodology

Proving a live kill-switch fires under a genuinely dangerous broadcast presents a real ethical constraint. The mechanism was proven with a designated substitute trigger standing in for

genuine extremist content — a deliberate choice that proves capture, transcription, evaluation, and stream termination all fire in sequence without manufacturing dangerous material to test it.

Verified behaviors

- **The live pipeline fires end to end during an active broadcast** — capture, transcription, evaluation, and stream termination in sequence. *Verified June 17, 2026.*
- **Live transcription works accurately in English and Spanish**, feeding moderation in near-real-time. *Verified June 17, 2026.*
- **Live content is evaluated within a short detection window**, not after a meaningful delay. *Verified June 17, 2026.*

Honest limitations

Real-time moderation operates within latency limits — screening a live stream is by nature a race. Consequential stream actions bias toward human escalation, not automation alone.