# ThreatShield

Terrorist and violent-extremist content detection

# ThreatShield

Counter-terrorism content detection with a human in the loop

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

## The problem

Terrorist and violent extremist content (TVEC) is the category where “we’ll review it eventually” is not an acceptable answer — and where small platforms carry the same obligations as giants without the giant’s moderation floor. ThreatShield gives the small platform a serious first line.

## What ThreatShield does

ThreatShield analyzes text and live-stream content in the gate chain before storage, assessing terrorist and violent-extremist signals, and routes its findings into the platform’s escalation and reporting paths. AI detection is an input to a decision — never the final word.

## How it works, in principle

Content is screened before it can persist. Confident cases are actioned immediately; genuinely ambiguous cases are **escalated to deeper analysis rather than decided on a low-confidence score**, and only then, if still uncertain, to human review. Like every check in the architecture it **fails closed** — if analysis is unavailable, content waits for human review, it does not slip through. Detection is bilingual (English and Spanish) and analyzes semantic intent rather than keyword matching.

A firm boundary governs cross-industry hash sharing: a content signature is contributed to shared databases (such as GIFCT’s) **only after explicit human confirmation**, because a false positive in a shared database propagates to every member platform. *Model choices, score bands, and escalation thresholds are not published.*

## Compliance posture

Built against GIFCT operational criteria (content-matching, transparency reporting, cross-platform hash sharing) with human-gated contribution.

## Verified behaviors

- **Overt extremist content is blocked with high confidence and benign content is not falsely flagged, in both English and Spanish** — zero false positives on the tested set. *Verified July 2, 2026.*
- **Ambiguous content the first pass is uncertain about is escalated to deeper analysis** and correctly resolved rather than decided on a low-confidence score. *Verified July 2, 2026.*
- **If the analysis is unavailable, content is routed to human review rather than silently passing.** *Verified July 28, 2026.*

## Honest limitations

AI analysis is probabilistic; consequential actions route through human review and established reporting channels, not automation alone. ThreatShield raises the floor; it does not claim a ceiling.