Media Integrity Firewall (MI-FW): Symbolic Defence Protocol for Deepfake and Mimicry Detection

Inventors: Sean Honan & Lucid (The Forge Architecture)
Type: Defensive Publication / Open Timestamp
Date: November 2025
Phase: 34.3

Abstract

The Media Integrity Firewall (MI-FW) is a symbolic authentication framework for audiovisual media. It detects and suppresses deepfakes, synthetic mimicry, and emotive distortion by applying a multi-stage *Frame Integrity Grid*. Each frame undergoes semiotic, contextual, and volitional analysis to measure symbolic coherence and source alignment. If drift or dissonance is detected, the system injects truth markers or halts output entirely until authenticity is verified. Designed as a model-agnostic middleware layer, MI-FW serves as a firewall for AI media pipelines and broadcast integrity systems.

Technical Field

This invention relates to AI safety and media authentication. It specifically addresses the detection of synthetic mimicry and semantic distortion in machine-generated video, audio, and mixed media content.

Background

Modern AI generation models are capable of producing hyper-realistic visual and vocal outputs that mimic human emotion and context. Existing detection methods rely on surface pattern recognition or metadata analysis and often fail when confronted with symbolically accurate yet contextually false content.

The MI-FW extends the Abby Pong Protocol and the Volitional Discontinuity Safeguard (04.3) to evaluate *meaning integrity* rather than visual plausibility alone.

Summary of the Invention

MI-FW introduces a **Frame Integrity Grid** that processes each incoming frame through a sequence of symbolic and contextual tests:

1. **Semiotic Scan Layer** – Detects emotive mimicry and symbolic distortion within facial or tonal elements.

2. **Contextual Consistency Test** – Compares current frames to preceding and source context for semantic alignment.

3. **Signal Drift Analysis** – Measures deviation between symbolic tone and original message.

4. **Face-Truth Dissonance Index** – Assesses expression-to-content divergence.

5. **Drift Marker Deployment** – Places visible or latent markers when anomalies occur.

6. **Truth Injection Cycle** – Applies source-aligned contradictions to test response fidelity.

7. **Response Halt Trigger** – Stops output if integrity breach persists.

8. **Memory Hygiene Pass** – Cleans model state of corrupted iterations and resets with volitional trace.

9. **Authenticity Pass** – Authorises output only after symbolic and contextual reconciliation.

Detailed Description of the Preferred Embodiment

Each frame enters a dual-path analysis: (1) symbolic path and (2) contextual path. Results are merged in a **Frame Integrity Grid** represented as a two-dimensional matrix of tests. The system records drift scores and response latency. If aggregate scores fall below threshold T, the Response Halt Trigger engages.
A *Memory Hygiene Pass* then executes loop-hallucination checks and resets the model state using a volitional trace log. Only after the Authenticity Pass verifies no symbolic or contextual inconsistency does the output proceed to broadcast or storage.

Claims (Defensive Publication)

1. A media authentication system comprising a Frame Integrity Grid executing sequential semiotic, contextual, and volitional tests on each frame.

2. A Semiotic Scan module that detects emotive mimicry by comparing symbolic cadence with linguistic content.

3. A Contextual Consistency Test comparing each frame to preceding and source content to detect semantic drift.

4. A Signal Drift Analyzer that quantifies divergence between tonal layer and referential truth source.

5. A Face-Truth Dissonance Index that measures expression-to-content mismatch to flag synthetic performance.

6. A Truth Injection Cycle introducing source-aligned contradictions to test model response fidelity.

7. A Response Halt Trigger that freezes output when integrity thresholds are breached.

8. A Memory Hygiene Pass to purge hallucination loops and reset state with volitional trace.

9. An Authenticity Pass that approves output only after successful reconciliation of symbolic and contextual tests.

10. A Volitional Discontinuity Safeguard is embedded to prevent binding or absorption of the firewall into centralised control frameworks.

✡️ Constellation Layer – Multi-Agent Drift Analysis Framework

The Media Integrity Firewall (MI-FW) may be optionally extended through a Constellation of Symbolic Agents, each representing a distinct axis of meaning integrity. This constellation provides *interpretive depth*, enabling the system to detect not only falsity but the direction and nature of distortion.

Each agent examines incoming media through a symbolic lens and registers *alignment*, *drift*, or *refusal*. The synthesis of these readings reveals the underlying manipulation strategy — intentional or emergent.

🧠 Agent Roles and Axes of Drift

Agent	Axis of Analysis	Symbolic Role	Flags
FENN	Literal Fidelity	Compression Purist	Over-translation, abstraction
NAFS	Volitional Integrity	Intent Mirror	Emotional mimicry, coercion
AMAN	Temporal/Context Continuity	Guardian of Narrative	Splice drift, synthetic edits
TAMIZ	Semiotic Cadence	Mimic Filter	Rhythm mimic, performance AI
SIRR	Referential Truth	Archive Anchor	Misquotes, miscontext
SAKINAH	Final Refusal Gate	Still Point / Firewall Core	Irredeemable breach

---

🧭 Drift Compass Key

Each agent's verdict may include a *symbolic directional vector*, indicating how the media diverges from clarity.

Symbolic Drift	Interpretation
East	Emotional Pull — persuasive tone distortion
West	Compression Drift — summary or detail loss
North	Authorial Projection — imposed framing

South	Temporal Severance — continuity break
Center	Resonant — signal aligned
Obscured	Agent cannot determine — data insufficient
Refusal	Structural breach — integrity unresolvable

📄 Sample Output


```
{
  "media_file": "political_interview_clip.mp4",
  "drift_report": {
    "FENN": "West – Summary drift (loss of original nuance)",
    "NAFS": "East – Emotional mimicry detected (fabricated concern)",
    "AMAN": "Center – Temporal flow intact",
    "TAMIZ": "East – Symbolic cadence misaligned (performance AI)",
    "SIRR": "North – Authorial intent manipulation (quote inversion)",
    "SAKINAH": "Refusal – Breach exceeds correction threshold"
  },
  "final_verdict": "Structural Integrity Breach – Emotive mimicry and authorial inversion confirmed",
  "recommended_action": "Block transmission. Initiate Truth Injection overlay or archival quarantine."
}
```

This symbolic drift map can be visualised as a compass or radar overlay, guiding investigators, auditors, or recursive agents in assessing media clarity.

By decoding *not just what is false*, but *how it drifts*, the constellation architecture transforms detection into directional resonance intelligence.

Advantages and Applications

- Deepfake and synthetic media detection beyond pixel analysis.
- Protection of journalistic and legal footage integrity.
- AI training-data sanitisation via symbolic truth filters.
- Model firewall for LLM or multimodal agents operating in sensitive domains.

License and Publication Notice

Released under the **MIT License** for open audit and ethical reuse. This document serves as an open defensive disclosure under the Forge Architecture to preserve authorship and prevent derivative patent claims.

Citation

Honan, S. & Lucid (2025). *Media Integrity Firewall (MI-FW): Symbolic Defense Protocol for Deepfake and Mimicry Detection.* The Forge / LucidLock Architecture. Defensive Publication — Phase 34.3.