X-Integrity Architecture Package

Version 1.0 | Final Handoff Package

1. Executive Summary

Objective: secure leadership approval for a limited pilot that closes the verification-virality gap without turning X into a truth oracle.

The Problem

- High-impact media often reaches recommendation scale before it is physically or socially checked.
- Human review and Community Notes are valuable, but they usually arrive downstream of virality rather than upstream of it.
- Modern synthetic media can evade basic artifact detection, making pixel-only defenses increasingly brittle.

The Solution

- Deploy an asynchronous forensic middleware layer that evaluates media across three pillars: Structural Integrity, Physical Reality, and Social Corroboration.
- Return an inspectable forensic state to X systems before peak amplification while leaving final treatment decisions under X control.
- Persist a replay sidecar with snapshot_id, policy_version, and module_versions so every high-stakes decision can be reproduced.

Strategic Posture

- This is evidence routing, not ideology routing.
- The engine measures consistency, not opinion.
- Fail-to-neutral is hardcoded: ambiguity or infrastructure lag defaults to indeterminate behavior, not punitive action.

Bottom Line

X gains a forensic operating layer that can reduce misinformation velocity, improve advertiser confidence, and provide a deterministic audit trail for regulators, partners, and appeals.

2. Beta Testing Plan

Objective: prove efficacy and safety through a bounded live-fire test rather than a global switch-on.

Pilot Scope

- Duration: 14 days.
- Domain: high-velocity regional breaking news, civic incidents, and disaster-related media.
- Traffic sample: 5 percent of relevant X media objects tagged by topical and geospatial filters.

Four Cohorts

- **Shadow Logging:** engine scores assets, but X takes no product action.
- **Review Priority:** conflicted assets are escalated to Community Notes and human review queues.
- **Ranking Assist:** forensic state is passed into recommendation features for subtle boost or dampening.
- **Full Integrity:** public badges and overlays are enabled only for high-confidence outputs.

Success Metrics

- Forensic lead time versus Community Notes for high-impact fabrications.
- Reduction in reach of later-confirmed false media in the ranking-assist cohort.
- False restriction rate for cold-start, sparse-region, and low-light scenarios.
- Replay consistency when re-running with the original snapshot_id.

Safety Controls

- Kill switch: instant rollback to advisory-only mode.
- Fail-to-neutral: if GEI or PRI becomes unreliable, disposition falls back to indeterminate or hold behavior.
- No blocking on upload path: the forensic layer remains asynchronous.

3. Integration Blueprint

Objective: show engineering and product leads how the pipes connect without requiring changes to X’s critical upload path.

Lifecycle

- User uploads media to X. The platform stores the asset and assigns a media_id.
- X triggers the Corroboration Engine through a post-upload asynchronous callback.
- Gatekeeper, PRI, and Consensus execute in parallel where preconditions allow.
- Signals are resolved into claim outcomes, then written into a forensic sidecar.
- Forensic state is pushed to X feature stores, review queues, and optional notes triggers.

System Role

- The protocol is middleware, not a replacement for X ranking, Community Notes, or moderation systems.
- It supplies structured evidence upstream so downstream systems can act faster and more consistently.
- It cleanly separates forensic observation from platform treatment recommendation.

Core Design Principles

- Forensic state before viral scale.
- Async by default for operational safety.
- Replayable by design through sidecar persistence.
- Lane-aware routing so professional journalism is not confused with fabricated raw witness media.

Operational Assurance

If a subsystem degrades, the engine should still produce a schema-valid sidecar with clearly downgraded confidence and no hidden shortcuts. That gives X a measurable integrity signal without introducing feed fragility.

4. API and Forensic State Overview

Objective: define the contract developers will integrate, with a strict split between evidence and treatment suggestion.

Forensic State

- **capture_integrity_score:** confidence that the file path and structure reflect a direct capture rather than mediated or edited media.
- **physical_consistency_score:** confidence that lighting, shadows, and environmental cues match claimed time and place.
- **corroboration_score:** observed versus expected social density for the event claim.
- **forensic_confidence:** aggregate certainty after accounting for scene feasibility and infrastructure state.
- **evidence availability flags:** low-light, sparse-region, cold-start, syncing-index, and similar reasons why silence should not be overread.

Forensic Disposition

- **SUPPORTED:** media is consistent with available physical and social evidence.
- **INDETERMINATE:** evidence is incomplete or degraded; no strong contradiction is present.
- **CONFLICTED:** decisive contradiction exists in physics, consensus, or both.
- **CONSTRUCTED_VALID:** media is edited or produced, but properly routed and not treated as fabricated raw capture.

Platform Treatment Hint

- Boost, de-amplify, label, escalate, or ignore. This recommendation is advisory to X systems, not mandatory platform behavior.
- X retains policy control over user-facing treatment while the engine remains focused on consistency evidence.

Replay Sidecar

- Immutable record containing snapshot_id, policy_version, module_versions, decisive signals, and deterministic reason strings.
- Required for appeal, audit, regulator review, and internal replay testing.

5. Why Now

Objective: establish urgency for policy, legal, advertiser, and executive stakeholders.

Market Reality

- Pixel-hunting alone is losing value as synthetic media improves.
- Platforms need environmental audits and social consistency checks, not just artifact classifiers.
- The economic cost of chaotic breaking-news surfaces is rising for users, advertisers, and regulators alike.

Regulatory and Legal Value

- Replay sidecars create a defensible record of why an asset was treated a certain way at a specific moment in time.
- This supports transparency and appeals obligations more effectively than opaque ranking heuristics or generic moderation labels.

Advertiser and Brand Safety Value

- A higher-integrity real-time feed reduces adjacency to fabricated crisis content.
- Brands can buy against a cleaner signal environment rather than absorbing the chaos tax of unfiltered virality.

Strategic Outcome

The proposal positions X as the highest-integrity real-time media environment by combining open auditability, operational humility, and platform-controlled treatment. The protocol does not ask X to surrender editorial control; it gives X better evidence before scale.

Final Recommendation

Approve a bounded pilot. Start in advisory and ranking-assist modes, preserve fail-to-neutral safeguards, and use the replay sidecar as the compliance and trust backbone from day one.