Nick Clark · Published April 25, 2026 · PDF

Forbidden-Content Blocking at Upload and Generation Time: Pre-Release Exclusion Against Signed Policy

Trust-and-safety teams and AI-generation platforms still block forbidden content after it has already been written to storage, returned from an endpoint, or shown to a viewer, so the impermissible artifact exists, is logged, and is exposed before any classifier flags it. This application moves the decision to the moment before a candidate is committed or released, evaluating it against a versioned, signed policy object and a governed exclusion corpus and rendering structurally impermissible content non-committable. It is built on Content Anchoring, the home inventive step disclosed in PCT International Application No. PCT/US26/28630.

What This Application Specifies

This application specifies forbidden-content blocking that happens before content is committed, applied to the trust-and-safety upload pipeline and to AI-generation guardrails. The home inventive step is the pre-release admissibility engine of the Content Anchoring invention disclosed in PCT International Application No. PCT/US26/28630, which interposes an admissibility evaluation between a candidate content artifact and its commitment. A commitment is any irreversible or externally visible side effect of a content event: public release, customer delivery, an endpoint return, a marketplace publication, training-data admission, or a cross-platform

provenance anchor. In the trust-and-safety setting the commitment boundary is the upload write or the generation return. The engine governs at that boundary rather than scanning a store after the fact.

The candidate enters the engine and is routed through two parallel tracks before a single commitment gate. The first track is a policy object evaluator that receives a versioned, cryptographically signed, machine-evaluable policy object and produces an admissibility decision. The second track is a structural similarity evaluator that computes cosine similarity between the candidate's multi-axis variance vector and the variance vectors of reference artifacts indexed in a governed exclusion corpus, and feeds that score to a forbidden-content exclusion layer that compares it against a policy-declared threshold. If both tracks confirm admissibility, the gate permits a committed artifact. If either track fails, a rejection handler emits a regeneration signal or an escalation signal in place of a commitment.

The variance vector is derived from the internal structure of the artifact itself. Per the disclosure, the multi-axis variance vector comprises a first axis encoding cross-scale energy distribution, a second axis encoding cross-scale frequency compaction, and a third axis encoding structural phase persistence from gradient orientation distribution, and it is operable across raster images, audio waveforms, textual documents, video frames, and binary objects normalized to a scalar field. The exclusion corpus is the same slope-band-indexed anchor network used throughout the platform, whose entries are registered under signed corpus policy objects that declare admissibility scope, exclusion classes, and similarity tolerance thresholds. A candidate whose variance vector falls within a configured proximity of any exclusion-corpus entry is rendered non-committable prior to release.

Why It Matters

Platform trust-and-safety and generative guardrails today share a structural defect that the disclosure names directly: conventional generative systems evaluate admissibility through post-generation moderation filters applied after an output artifact has been produced, which cannot prevent impermissible content from existing as an internal artifact and do not provide reproducible, auditable decisions verifiable from versioned policy records. The same is true of an upload pipeline that accepts a file, persists it, and only then runs a classifier. The forbidden artifact has already been written, replicated to a cache, and possibly served before the verdict returns. Every post-hoc architecture is racing its own storage layer.

Moving the decision before commitment removes the race. The disclosure states the property plainly: governance prevents forbidden content from existing as released media rather than filtering it after exposure has already occurred. For a trust-and-safety operator this is the difference between blocking an upload and taking down a post that viewers, caches, and downstream syndication already saw. For an AI-generation provider it is the difference between refusing to return an output and recalling one that an API consumer already holds.

The second reason is auditability. Because the policy object is versioned and signed and the candidate is identified by its variance-derived unique identifier and structural signatures, an admissibility decision is reproducible: any authorized party can replay the evaluation against the same artifact identity and the same policy version and confirm the determination. This is structurally different from an opaque classifier whose score cannot be independently re-derived. A trust-and-safety team that must justify a block to a regulator, an appeals process, or an internal audit can show the governing policy version and the structural identity that triggered exclusion, rather than asserting that a model returned a number.

How It Composes With the Domain

In an upload pipeline the integration point is the moment between receiving the candidate and writing it to durable storage. The disclosure describes a client-side execution architecture in which canonical resizing, grayscale conversion, orientation canonicalization, multi-scale variance analysis, gradient-histogram computation, and the 320-bit unique-identifier hash all run in a standard browser execution context using the Canvas 2D API and standard JavaScript, without WebGL, WebAssembly, GPU compute, or a per-query inference service. The raw artifact does not leave the device during evaluation; only the computed unique identifier and the resulting decision are transmitted. For trust-and-safety this composes cleanly with data-minimization obligations that restrict transmission of personal media, because the block can be decided at the edge before the file is ever sent.

The exclusion comparison is bounded by the slope-band structure rather than by a full-corpus scan. A candidate's global variance assigns it to a variance band, and the structural similarity evaluator narrows the comparison to the relevant band cluster instead of scanning every reference. The disclosure further provides a locally cached exclusion-corpus fragment: a slope-band-filtered subset of the governed corpus pre-fetched for the content categories the client is authorized to process, together with a locally stored signed policy object, both verifiable by their cryptographic signatures without a live connection at evaluation time. An upload checkpoint can therefore block known-forbidden classes even in disconnected or intermittently connected operation, with assurance that the corpus fragment and policy object are authentic and unmodified.

For AI-generation guardrails the candidate is the model output, and the gate sits before the output is returned. The disclosure describes evaluating a generated artifact through the same extraction pipeline without regard to the model that produced it, and routing the structural-similarity result and policy-object verdict to the commitment gate. On failure the rejection handler can issue a regeneration signal that directs the producer to

regenerate under modified generation constraints, which maps directly to a guardrail that re-rolls a generation rather than emitting a refusal, or an escalation signal that routes the determination to an authorized override authority named in the policy object. Real-time and streaming generation is supported by a sliding-window mode: when the cosine similarity between a window-level unique identifier and a registered reference exceeds the policy-declared threshold, the system generates a real-time match event that may trigger blocking of unauthorized retransmission or invoke the pre-release admissibility engine.

The policy object is the unit of configuration that the domain plugs into. It declares typed category constraints, jurisdictional scopes, override authorities, similarity tolerance thresholds, and escalation paths. A trust-and-safety organization expresses its forbidden classes, its regional rules, and its appeal routing as a signed policy version; when the rules change, a new signed version is issued and decisions made under it remain replayable against that version.

What This Enables

A platform can block at the boundary instead of moderating after exposure, so a forbidden upload is never written and a forbidden generation is never returned. It can do this at the scale of real-time content production, because similarity is evaluated over variance-derived unique identifiers rather than through GPU inference or a centralized embedding index, with no per-query compute cost proportional to corpus size. The resolution protocol supports bulk resolution, in which a client submits a batch of candidate unique identifiers across variance bands in one request and the anchor network routes each to its band cluster in parallel; the disclosure frames this for upload pipelines and content-moderation systems resolving large volumes per second at the latency of a single amortized round-trip.

The decision is portable and reproducible. Because nothing is embedded in the artifact and no enrollment is required, the same variance vector computed at upload, at generation, and at a later audit yields the same identity, and the admissibility decision can be replayed against the recorded policy version. A platform can keep a defensible record that a given block was made under a given signed policy, and an appeals reviewer can re-execute it. Consultation and governance records extend this to provenance: generation events that consult reference artifacts are deterministically logged, and training-corpus admission is recorded as a verifiable lineage, so a generation guardrail can sit inside a broader chain that ties outputs back to governed inputs.

Boundary Conditions

The block is only as good as the policy object and the exclusion corpus behind it. The disclosure governs structure, not semantics: it does not declare which content categories are forbidden, which cryptographic primitives sign a policy object, or which escalation endpoints a rejection reaches. Those are deployment choices, and an operator that populates the exclusion corpus poorly or sets the similarity threshold loosely will under-block regardless of the architecture.

Exclusion here is structural proximity in variance space, not classification of meaning. It is strong against re-encodes, format conversions, rescales, and lossy compression of a known-forbidden reference, because the variance vector is designed to be stable under those controlled transformations while diverging predictably under content-altering edits. It is not a substitute for a semantic policy decision about novel content that resembles nothing in the corpus; for that, the policy-object track and human escalation paths carry the load. The disclosure also notes that structurally unanchored artifacts, such as freshly synthesized outputs with no registered lineage, are not necessarily impermissible; they trigger heightened scrutiny under policy objects that govern synthetic content rather than an automatic block.

Disconnected operation depends on the freshness of the cached corpus fragment and policy version. A device evaluating against a stale fragment blocks against an older exclusion set; the signatures guarantee authenticity, not currency. The stable-under-transformation property holds within defined thresholds, so an adversary who mutates a forbidden reference far enough in variance space can move it outside a configured proximity radius, which is why the disclosure pairs exclusion with similarity-tolerance thresholds and escalation rather than treating a single radius as complete.

Disclosure Scope

The technology described here is disclosed in PCT International Application No. PCT/US26/28630: the pre-release admissibility engine and its evaluation at the commitment boundary; the policy object evaluator producing a reproducible decision from a versioned signed policy object; the structural similarity evaluator computing cosine similarity between a candidate's variance vector and reference artifacts in a governed corpus; the forbidden-content exclusion layer that renders a candidate non-committable when its variance vector falls within a configured proximity of an exclusion-corpus entry; the commitment gate and the rejection handler emitting regeneration or escalation signals; the client-side execution architecture and locally cached, signed exclusion-corpus fragment; and the bulk resolution protocol. Every claim above about what the system does traces to that filing.

The trust-and-safety and AI-generation framing is external enabling context, not part of the disclosed invention. Specific forbidden-content categories, regional or jurisdictional rule sets, appeal and escalation workflows, regulator-facing reporting obligations, and the commercial moderation products a platform might replace are deployment and domain facts described here only to show a faithful implementation. The disclosure constrains the structural properties: the evaluation must be interposed before commitment, the policy object must be versioned and signed so decisions are reproducible, similarity must be evaluated over variance-derived unique identifiers rather than embedded markers or opaque classifier outputs, and impermissible content

must be rendered non-committable rather than filtered after release. A system that moderates only after upload or after a generation is returned falls outside the disclosure.

Content Anchoring (/content-anchoring)

All 40 steps → (/inventive-steps)

Computable identity for media. Provenance from structural variance.

PCT/US26/28630 (/patents/pct-us26-28630)

PRIMARY TECHNICAL DISCLOSURE

- Content Anchoring: Computable Identity for Media That Changes (/articles/content-anchoring-computable-identity-for-media-that-changes)

SECONDARY TECHNICAL

- Multi-Axis Variance Vector Extraction: Nine Dimensions of Structural Content Identity (/articles/content-anchoring/variance-vector)
- Quadrant Decomposition: Spatial Sub-Region Fingerprinting for Partial Similarity Detection (/articles/content-anchoring/quadrant-decomposition)
- 320-Bit UID Construction: Multi-Segment Hashing for Negligible Collision Probability (/articles/content-anchoring/uid-construction)
- Structure Signature: Background-Invariant Matching Through Gradient-Only Descriptors (/articles/content-anchoring/structure-signature)
- Constellation Signature: Geometry-Invariant Matching Across Crop, Scale, and Occlusion (/articles/content-anchoring/constellation-signature)
- Five-Band Variance Classification: Content Routing by Structural Complexity (/articles/content-anchoring/variance-classification)
- Variance Saturation-Governed Cache Eviction: UID Density Replacing Static TTL (/articles/content-anchoring/cache-eviction)
- Multi-Root Composite Lineage Graphs: Provenance Through Variance Vector Similarity (/articles/content-anchoring/composite-lineage)
- Multi-Modal Content Identity: Unified Pipeline Across Image, Audio, Text, and Video (/articles/content-anchoring/multi-modal-identity)

- Rights-Grade Pre-Release Admissibility: Policy Evaluation Before Content Commitment (/articles/content-anchoring/pre-release-admissibility)
- Training Corpus Governance: Verifiable Lineage From Training Data to Model (/articles/content-anchoring/training-corpus-governance)
- Consultation Event Logging: Deterministic Records of Every Generation Reference (/articles/content-anchoring/consultation-logging)
- Model Output Provenance Fingerprint: Structural Proximity Without Model Access (/articles/content-anchoring/output-provenance)
- Creator Attribution and Compensation Routing: Payment From Consultation Lineage (/articles/content-anchoring/creator-attribution)
- Adversarial Robustness and Deepfake Detection: Content Identity as Detection Substrate (/articles/content-anchoring/adversarial-robustness)
- Client-Side Execution Architecture: Privacy-Preserving Variance Computation on Device (/articles/content-anchoring/client-side-execution)
- UID Resolution Query Protocol: Distributed Lookup Across Anchor Node Networks (/articles/content-anchoring/uid-resolution)
- Orientation Canonicalization: Rotation-Invariant Processing Through Gradient Normalization (/articles/content-anchoring/orientation-canonicalization)
- Cross-Band Resolution Pathfinding: Traversal Between Variance Bands Under Mutation (/articles/content-anchoring/cross-band-resolution)
- Identity by Position: Media as a Third Navigable Space (/articles/content-anchoring/identity-by-position)

APPLICATIONS · GENERAL

- **Forbidden-Content Blocking at Upload and Generation Time: Pre-Release Exclusion Against Signed Policy (/articles/content-anchoring/forbidden-content-blocking)**
- Structural Provenance for Software Supply Chains: Binary and Firmware Identity Independent of SBOM Metadata (/articles/content-anchoring/software-supply-chain-provenance)
- Rights-Grade Generative AI: How to Pay Creators, Exclude Forbidden Content, and Prevent Infringement Before Release (/articles/content-anchoring/rights-grade-generative-ai)
- Deepfake Detection by Structural Provenance: Verifying Synthetic Media Without Watermarks (/articles/content-anchoring/deepfake-provenance)
- Creator Economy Attribution Without Platform Intermediaries (/articles/content-anchoring/creator-attribution-economy)
- Verifying Source Photos and Video in the Newsroom: Content Anchoring for Journalism (/articles/content-anchoring/journalism-verification)

- Detecting Image Manipulation and Proving Figure Provenance in Research Publications (/articles/content-anchoring/academic-research-integrity)
- Content Anchoring for Legal Evidence Chains (/articles/content-anchoring/legal-evidence-chain)
- Content Anchoring for Insurance Claims Evidence (/articles/content-anchoring/insurance-claims-evidence)
- Content Anchoring for Real Estate Documentation (/articles/content-anchoring/real-estate-documentation)
- Art Authentication and Provenance Verification with Content Anchoring (/articles/content-anchoring/art-authentication)

APPLICATIONS · SPECIFIC

- C2PA Attaches Provenance to Content. The Content Itself Has No Identity. (/articles/content-anchoring/c2pa)
- Google SynthID Watermarks AI Output. Watermarks Are Not Identity. (/articles/content-anchoring/google-synthid)
- Shutterstock Tracks Licensed Media. The Media Itself Cannot Prove Its Own Identity. (/articles/content-anchoring/shutterstock)
- Spotify Tracks Every Stream. The Music Itself Has No Computable Identity. (/articles/content-anchoring/spotify)
- Getty Images Built the World's Largest Licensed Image Library. Image Identity Still Depends on Metadata. (/articles/content-anchoring/getty-images)
- Adobe Stock Integrates Licensed Content Into Creative Workflows. Content Identity Is Still External. (/articles/content-anchoring/adobe-stock)
- YouTube Content ID Matches Audio and Video. The Content Has No Intrinsic Identity. (/articles/content-anchoring/youtube-content-id)
- Audible Magic Identifies Audio Content. The Audio Has No Self-Identifying Properties. (/articles/content-anchoring/audible-magic)
- Digimarc Embeds Invisible Watermarks. The Watermark Is Added, Not Intrinsic. (/articles/content-anchoring/digimarc)
- Irdeto Protects Digital Content Through DRM. The Protection Is Applied, Not Intrinsic. (/articles/content-anchoring/irdeto)

Content Anchoring overview → (/content-anchoring)