


# Algorithmic Proximate Harm Detection: A Forensic Framework for Causal Accountability in Platform Design

Bee Rosa Davis 

NASA / Independent Researcher

Sacramento, CA

bee\_davis@alumni.brown.edu

**Abstract**—Algorithmically mediated harm is increasingly recognized as a structural failure of digital platform design, yet prevailing accountability frameworks remain limited to group-level disparity audits, anecdotal reports, or post hoc moderation reviews. These approaches lack the traceable specificity required for user-level redress, forensic analysis, or legal inquiry. This paper introduces the *Algorithmic Proximate Harm Detection (APHD)* framework—a modular, constraint-aware scoring system for identifying and decomposing patterns of platform-amplified abuse at the individual user level.

APHD formalizes a nine-variable scoring framework structured around a triadic model of harm attribution—exposure architecture, governance failure, and behavioral reinforcement. The variables include graph centrality, temporal proximity, recurrence, platform negligence, exposure delta, behavioral reinforcement, silence suppression, friction, and design bias. These are synthesized into two interpretable outputs: the *Proximate Harm Score (PHS)* and the *Systemic Harm Index (SHI)*. APHD employs a bounded optimization procedure (AMSBA) and SHAP-based decomposition to ensure transparency, explainability, and forensic auditability.

Rather than asserting experimental causality, APHD provides a structured method for surfacing and analyzing harm pathways consistent with legal standards of traceability and foreseeability. We demonstrate the framework’s logic through a simulated harm graph scenario and outline a validation roadmap encompassing red-team simulation, pilot deployment, and survivor-aligned expert review. APHD is designed to support—not supplant—legal and investigative processes by rendering digital harm legible, auditable, and explainable under conditions of partial observability.

**Index Terms**—Algorithmic harm, forensic causation, graph theory, AMSBA, SHI, platform accountability, recommender systems, APHD

## I. INTRODUCTION

Digital platforms now operate at a speed, scale, and algorithmic depth that outpace traditional regulatory frameworks. Optimization engines—designed to maximize engagement—routinely surface content and users without structural constraints for safety, dignity, or harm prevention. For structurally marginalized users, especially those navigating queer or racialized digital spaces, this logic often produces repeated exposure to aggressors, reappearance of blocked users, and silencing through inaccessible or adversarial moderation workflows.

While fairness audits and group disparity analyses have improved public understanding of algorithmic bias, they remain limited in scope: retrospective, aggregate, and rarely usable in legal or forensic contexts. These tools do not address how harm unfolds for a specific user over time—through recommender exposure, moderation failure, and systemic reinforcement. Nor do they support the traceability, decomposability, or evidentiary alignment required for harm scoring under constrained observability.

This paper introduces the *Algorithmic Proximate Harm Detection (APHD)* framework: a modular, explainable system for detecting and scoring user-level exposure to platform-mediated harm. APHD is grounded in a triadic theory of algorithmic harm that identifies three causal mechanisms: (1) exposure architecture, (2) governance failure, and (3) behavioral reinforcement. These are operationalized through nine traceable variables derived from interaction logs, platform configurations, and visibility telemetry. The variables are synthesized into two outputs: the *Proximate Harm Score (PHS)*, capturing acute, user-level harm patterns, and the *Systemic Harm Index (SHI)*, measuring design-level amplification and negligence.

Rather than asserting counterfactual causality, APHD models traceable harm pathways aligned with forensic standards of proximate cause: foreseeability, failure to intervene, and repeated exposure under platform control. Scores are optimized using a constrained metaheuristic (AMSBA) and decomposed via SHAP to ensure transparency, auditability, and alignment with explainable AI standards.

This framework is grounded in lived experience. As a Black transgender woman using queer dating platforms, I was repeatedly surfaced to known abusers—even after reporting them. In several cases, reporting increased my algorithmic visibility, suggesting structural reward for high-engagement harm. APHD emerges from this context—not to simulate harm hypothetically, but to render it computationally visible and investigatively actionable.

This paper makes four contributions:

- A triadic theory of algorithmic harm (exposure, governance, reinforcement), operationalized through nine log-traceable variables;

- A modular, decomposable scoring architecture (PHS and SHI) for modeling user-level harm and system-level negligence;
- A bounded optimization routine (AMSBA) integrated with SHAP constraints to ensure forensic interpretability;
- A synthetic case simulation and validation roadmap outlining APHD’s utility in audit, policy, and survivor-aligned review.

By modeling harm not merely as statistical disparity but as a traceable sequence of design, exposure, and inaction, APHD reframes digital trauma as both a public health emergency and a forensic failure—one that can, and must, be measured.

## II. RELATED WORK

Research on algorithmic harm has expanded across domains including fairness in machine learning, content moderation, platform governance, and explainable AI. However, few frameworks offer traceable, user-level modeling of harm with the specificity and granularity required for litigation, regulatory review, or forensic analysis. The *Algorithmic Proximate Harm Detection (APHD)* framework addresses this gap by synthesizing insights from four key research areas while contributing a distinct model of constraint-aware, graph-based harm inference.

### A. Bias Auditing and Fairness Metrics

Foundational work in algorithmic fairness has focused on detecting and correcting population-level disparities. Landmark contributions include facial recognition bias audits (Buolamwini & Gebru, 2018), disparate impact theory (Barocas & Selbst, 2016), and formal fairness constraints (Hardt et al., 2016). These approaches are vital for diagnosing systemic inequity but are typically retrospective, non-causal, and limited in their applicability to individualized harm contexts. They do not account for recursive exposure, design-layer risk propagation, or the evidentiary needs of forensic review.

### B. Content Moderation and Platform Infrastructure

Scholars such as Gillespie (2018) and Klonick (2017) have framed content moderation as a core function of platform governance. More recent work by Paresh et al. (2023) explores user friction, reporting abandonment, and the opacity of enforcement pathways. These analyses expose design vulnerabilities but stop short of providing a formalized mechanism for modeling harm recurrence or scoring traceable exposure failures. APHD builds on this literature by translating governance friction into quantifiable forensic variables.

### C. Explainability and Legal Interpretability

Techniques such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) have advanced post hoc model interpretability. Raji et al. (2020) extended these methods toward organizational audits. However, most interpretability approaches remain diagnostic—they expose model mechanics but do not produce evidence-ready decompositions. APHD repurposes SHAP not as an explanation aid but as a forensic

constraint: ensuring that each score component aligns with observable platform behavior and supports transparency under evidentiary review.

### D. Harm Classification and Causal Modeling

Recent work by Ravi and Yuan (2024) introduces behavioral taxonomies for classifying online harm, including escalation and grooming dynamics. These classification models surface important signals but do not formalize a traceable path from system design to experienced harm. Causal inference frameworks (e.g., Pearl, 2009; Glymour et al., 2016) offer theoretical models of causality but often rely on assumptions (e.g., do-calculus, instrumentality) that are difficult to satisfy in platform contexts. APHD operates in observational environments with incomplete data and encodes harm inference through recurrent signal alignment, topological features, and logged user-platform interaction traces.

### E. Distinct Contribution of APHD

Unlike prior work, APHD models user-specific harm as a dynamic, graph-structured, and system-amplified sequence. It scores not only exposure events but also the compounding effect of recommender pathways, moderation gaps, and platform design defaults. It is the first forensic scoring framework designed to support legal and regulatory accountability by producing log-traceable, SHAP-audited, and modular harm scores. APHD does not claim to assert ground-truth causality—it offers traceable, explainable scoring that aligns with the evidentiary standards required for harm attribution in real-world institutional contexts.

## III. THEORETICAL FRAMEWORK: A NETWORKED THEORY OF ALGORITHMIC HARM

The APHD framework is grounded in a theory of algorithmic harm as a networked, platform-mediated failure of safety, visibility, and accountability. It defines harm not as a singular event, but as an emergent property of multi-factor amplification—driven by three categories of system behavior:

- 1) **Exposure architecture** — the structures that determine who is seen, resurfaced, or targeted;
- 2) **Governance failure** — the latency, absence, or breakdown of moderation or intervention;
- 3) **Behavioral reinforcement** — the feedback loops that elevate harmful behavior or suppress vulnerable users.

From this triadic model, APHD defines nine forensic variables representing the minimal set necessary to detect, explain, and score harm within observable system telemetry. These variables are not arbitrary; they reflect distinct harm mechanisms repeatedly observed in platform audits, survivor testimony, and content moderation research.

- **Exposure mechanisms:**  $C_i$  (graph centrality),  $T_i$  (temporal proximity),  $\Delta E_i$  (exposure delta);
- **Governance signals:**  $P_i$  (platform negligence),  $S_i$  (silence suppression),  $F_i$  (friction),  $D_i$  (design bias);
- **Reinforcement patterns:**  $R_i$  (recurrence),  $B_i$  (behavioral reinforcement).

Other candidate variables—such as linguistic toxicity, bystander intervention, or reputational signals—were excluded either because (1) they cannot be consistently observed in platform-side logs, or (2) they are encoded in learned embeddings used by behavioral modules (e.g., SignatureProfiler), which detect latent language and graph signatures without requiring explicit symbolic representation.

APHD does not claim these nine variables capture the entirety of harm. It claims they form a forensic core: a constrained, explainable foundation for modeling traceable harm pathways under evidentiary constraints. Like other forensic frameworks—from financial risk scoring to forensic linguistics—these variables are derived from known harm dynamics and refined through constraint-aware scoring and validation.

This theory does not attempt to replace clinical, sociological, or cultural understandings of harm. It complements them by offering a traceable, log-anchored, and audit-compatible way to detect harm amplification when direct observation is incomplete or institutionally unavailable.

*a) Why Nine Variables?:* The APHD variable set reflects a deliberate design tradeoff between expressive power and forensic tractability. These nine dimensions were selected because they (a) map to observable platform behaviors and system traces, (b) decompose cleanly under SHAP for interpretability, and (c) align with the triadic theory of exposure, governance, and reinforcement. This set is not exhaustive—but it is sufficient for constructing structured harm inferences under partial observability. Future extensions may introduce additional variables, but this core is optimized for explainable, constraint-bounded forensic scoring.

#### IV. THE APHD FRAMEWORK

The *Algorithmic Proximate Harm Detection (APHD)* framework constructs a directed harm graph  $G = (V, E)$  to model the propagation, reinforcement, and governance failure patterns that characterize platform-mediated abuse. Nodes  $i \in V$  represent either users or critical system components implicated in harm trajectories, while edges encode time-stamped interactions, visibility flows, recommender referrals, or moderation outcomes. This graph-based abstraction allows both structural and temporal dynamics of harm to be rendered traceable and computationally explicit.

Each node is scored along nine forensic dimensions, selected to reflect distinct harm mechanisms derived from platform telemetry, behavioral signals, and governance metadata:

- $C_i$  (**Graph Centrality**): Measures the node’s embeddedness in exposure pathways using a composite of eigenvector, betweenness, and Fiedler centrality.
- $T_i$  (**Temporal Proximity**): Captures the recency and escalation of harmful interactions, giving greater weight to recent or intensifying events.
- $R_i$  (**Recurrence**): Quantifies reappearance of abuse patterns after block/report events—especially when linked to evasion behaviors.

- $P_i$  (**Platform Negligence**): Models inaction or delay in response to harm reports, including unacknowledged cases or missed moderation windows.
- $\Delta E_i$  (**Exposure Delta**): Measures post-report increases in visibility, flagging potential feedback loops or harm amplification.
- $B_i$  (**Behavioral Reinforcement**): Identifies whether the platform algorithmically boosted harmful behavior, such as surfacing users exhibiting high-risk patterns.
- $S_i$  (**Silence Suppression**): Detects disengagement or reporting dropout, often following institutional inaction or retraumatization.
- $F_i$  (**Friction**): Estimates the user effort required to report harm, including interface complexity, appeal limitations, or automated dismissals.
- $D_i$  (**Design Bias**): Encodes structural features (e.g., forced visibility or public-by-default settings) that elevate baseline exposure risk.

These variables are synthesized into two composite, SHAP-decomposable scores:

$$PHS_i = \alpha C_i + \beta T_i + \gamma R_i + \delta P_i + \epsilon \Delta E_i + \zeta B_i \quad (1)$$

$$SHI = \lambda_1 P_i + \lambda_2 \Delta E_i + \lambda_3 B_i + \lambda_4 S_i + \lambda_5 F_i + \lambda_6 D_i \quad (2)$$

The *Proximate Harm Score (PHS)* captures user-level harm trajectories—highlighting temporal escalation, recurrence, and structural vulnerability—while the *Systemic Harm Index (SHI)* quantifies broader failures in platform design and moderation logic.

Coefficient sets  $\Theta$  and  $\Lambda$  are optimized using a constrained metaheuristic (Adaptive Multi-Stage Bat Algorithm, AMSBA) to align scoring weights with platform-level harm traces and survivor-reported outcomes. Crucially, all weights are bounded and must pass a SHAP-based alignment constraint, ensuring that final scores remain auditable, interpretable, and legally reviewable.

Rather than asserting experimental causality, APHD identifies \*traceable harm pathways\* consistent with legal standards of proximate cause: foreseeability, repeated exposure, and failure to mitigate. The system supports forensic investigation by surfacing decomposable risk signals aligned with known harm dynamics. Every score can be explained, every variable audited, and every inference tied back to observable platform data. This enables APHD to operate as a structured, modular system for accountability—suitable for expert review, regulatory auditing, or survivor-led investigation.

While PHS and SHI are often correlated, they are not colinear: a user may experience high proximate harm even when system-level negligence scores low—particularly in isolated exposure edge cases.

#### V. MATHEMATICAL METHODS

We formalize algorithmic harm as a traceable process unfolding over a directed harm graph  $G = (V, E)$ , where nodes

$V$  represent users, moderation systems, or content features, and edges  $E$  encode directional exposure, interaction, or governance failure relationships. Edge weights reflect exposure strength, temporal adjacency, or ranking amplification, and are derived from logged user interactions, surfacing events, and moderation traces. This abstraction supports both structural and temporal harm inference.

For any node  $i \in V$ , APHD computes a vector of forensic features that represent exposure dynamics, recurrence behavior, and platform amplification. Scores are computed even in partially observable environments, with unpopulated variables excluded from aggregation and uncertainty flagged at the composite level.

While each scoring formula includes parameters (e.g.,  $\lambda$ ,  $\mu$ ,  $\theta_k$ ) that may initially appear subjective, these parameters are:

- **Bounded** within a restricted domain (e.g.,  $[0, 1]$  for influence weights),
- **Calibrated** using platform-specific logs, moderation targets, or empirically derived behavioral escalations,
- **Auditable**, with values exposed and documented for expert review or regulatory discovery.

Each feature score is normalized to  $[0, 1]$  before composition into  $PHS_i$  and  $SHI$ . Uncertainty due to missing data is explicitly tracked, and variables with unavailable input are omitted from composite score aggregation, with their absence noted in SHAP decomposition metadata.

Although APHD variables are modeled modularly to enable transparent decomposition, we do not assume strict statistical independence among them. Latent correlations—such as between platform negligence ( $P_i$ ) and reporting friction ( $F_i$ )—are implicitly captured through joint optimization over  $\Theta$  and  $\Lambda$ , which adjusts weights in response to correlated signal patterns in harm graphs.

To further mitigate redundancy and ensure decomposition integrity, we evaluate post-optimization multicollinearity using SHAP interaction values. This audit step identifies cases where two or more variables dominate through shared attribution, and flags them for weight rebalancing or reporting bias analysis. The framework thus remains sensitive to multivariate dynamics while preserving interpretability and legal auditability.

While APHD treats variables as modular for purposes of scoring transparency and SHAP decomposition, the optimization process captures latent interdependencies through multivariate calibration. Explicit covariance modeling is reserved for future ensemble extensions and is not assumed to be zero.

Future work may extend APHD with structured correlation-aware scoring ensembles, but current implementations rely on regularized joint optimization and SHAP-based post-hoc verification to balance fidelity, interpretability, and evidentiary clarity.

#### A. Graph Centrality ( $C_i$ )

Measures the structural role of node  $i$  in harm propagation:

$$C_i = w_1 \cdot EC_i + w_2 \cdot BC_i + w_3 \cdot FVC_i \quad (3)$$

Weights are constrained to  $w_1 + w_2 + w_3 = 1$  and tuned based on platform topology or past harm graph calibration. Graph structure is derived from a sliding window of interactions and system logs reflecting referral, resurfacing, and co-occurrence events.

#### B. Temporal Proximity ( $T_i$ )

Captures escalation and recency of harm events:

$$T_i = \left[ \sum_{k=1}^n w_k \cdot e^{-\lambda \Delta t_k} \right] \cdot \left( 1 + \theta_1 \cdot \frac{dI}{dt} \right) \quad (4)$$

$\lambda$  reflects a harm-type-specific temporal decay rate (e.g., slower for stalking, faster for harassment), and  $\theta_1$  encodes expected harm acceleration. These parameters are empirically derived or provided as policy inputs per platform and harm type.

$I(t)$  denotes a session-level cumulative harm index constructed from exposure intensity, engagement proximity, and interaction density. The derivative  $\frac{dI}{dt}$  models the rate of escalation in harm accumulation.

#### C. Recurrence ( $R_i$ )

$$R_i = \sum_{k=1}^n w_k \cdot f(\Delta t_k) \cdot (1 + \mu \cdot \mathbb{I}_{\text{evasion}_k}) \quad (5)$$

The  $\mu$  multiplier adjusts recurrence when evasion behaviors (e.g., banned user reappearance) are detected using identity linkage across IP/device/session embeddings. Conservative scoring is supported via  $\mu = 0$ .

#### D. Platform Negligence ( $P_i$ )

$$P_i = \frac{R_u}{R_t} + \theta_2 \tau_i + \theta_3 (1 - A_i) + \theta_4 H_p \quad (6)$$

Where  $R_u$  and  $R_t$  are counts of unresolved and total reports,  $\tau_i$  is moderation latency,  $A_i$  is binary acknowledgment, and  $H_p$  is harm severity (categorical).  $\theta_k$  weights are calibrated from internal moderation policies or industry norms.

#### E. Exposure Delta ( $\Delta E_i$ )

$$\Delta E_i = \frac{V_{\text{post}} - V_{\text{pre}}}{\max(V_{\text{pre}}, 1)} \cdot (1 + \phi \cdot \mathbb{I}_{\text{mitigation fail}}) \quad (7)$$

This score reflects relative exposure shift post-reporting. Positive  $\Delta E_i$  suggests failure to suppress visibility following harm escalation.

### F. Behavioral Reinforcement ( $B_i$ )

$$B_i = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\text{pattern match}_j} \cdot \mathbb{I}_{\text{boost}_j} \quad (8)$$

Each term reflects a SHAP-auditable logic rule: pattern match via SiRi, and confirmation of algorithmic amplification via surfacing traces.

### G. Silence Suppression ( $S_i$ )

$$S_i = \mathbb{I}_{\text{prior reports}} \cdot \left( \frac{\text{Expected} - \text{Observed}}{\text{Expected}} \right) \quad (9)$$

No parameters are learned; expected reporting frequency is estimated from cohort-specific baselines. SHAP output highlights this feature when reporting drop-off is anomalous.

### H. Friction ( $F_i$ )

$$F_i = \frac{\text{Clicks}}{\text{MinClick}} + \mathbb{I}_{\text{AutoDismiss}} + \mathbb{I}_{\text{NoAppeal}} \quad (10)$$

Captured directly from report path telemetry. Each binary term reflects known anti-user friction patterns in UI audits.

### I. Design Bias ( $D_i$ )

$$D_i = \sum_{j=1}^m d_j \cdot r_j \quad (11)$$

Where  $d_j$  is a feature flag (e.g., default public profile) and  $r_j$  is a platform-defined risk multiplier. All components are auditable via config snapshot or harm exposure audit logs.

### J. Coefficient Optimization via AMSBA

To calibrate score aggregation, we optimize weight vectors  $\Theta$  and  $\Lambda$  using an Adaptive Multi-Stage Bat Algorithm (AMSBA), selected for its ability to converge under interpretability constraints while preserving computational efficiency. We define:

$$\mathcal{L}(\Theta, \Lambda) = \sum_{i \in \mathcal{H}} \left| \hat{H}_i - H_i \right| + \Omega(\Theta, \Lambda) \quad (12)$$

where:

- $H_i$  is the observed harm indicator (e.g., disengagement, verified report, escalation);
- $\hat{H}_i = f(PHS_i, SHI_i)$  is the predicted harm score;
- $\Omega(\cdot)$  enforces boundedness, SHAP consistency, and non-negative feature attribution.

AMSBA proceeds through staged updates where each candidate “bat” iteratively improves its weight vector, and terminates when  $\mathcal{L}$  converges or stabilizes.

#### Constraints:

- $\sum \Theta = 1$  and  $\sum \Lambda = 1$  (soft normalization),

- $\Theta, \Lambda \in [0, 1]^n$  (bounded simplex),
- **SHAP alignment constraint:** To enforce evidentiary consistency, APHD requires that SHAP-derived feature attributions  $\phi_j$  maintain rank-order similarity to the optimized weights  $\Theta^*, \Lambda^*$  within a tolerance threshold  $\delta$ . This alignment is quantified using normalized Spearman rank correlation between the attribution vector and the coefficient vector for each score. If the correlation falls below  $\delta$ , a penalty is applied via the regularization term  $\Omega(\Theta, \Lambda)$ .

This constraint is enforced outside the primary gradient loop—after scoring, not during it—avoiding circular dependencies. Each candidate weight set is provisionally evaluated, its output decomposed with SHAP, and its alignment scored. Poorly aligned candidates are not discarded outright but penalized in the loss function, balancing stability with transparency. This ensures that scoring remains interpretable, legally defensible, and resilient to weight drift.

Normalized Spearman correlation is used instead of raw absolute difference to ensure invariance to scale and unit differences between coefficient sets and attribution vectors.

Final coefficients are publicly reportable and tied to platform configuration context. They are not interpreted in isolation, but decomposed via SHAP for expert explanation and dispute resolution.

### K. Graph Instantiation Protocol

Given a raw platform log dataset  $\mathcal{L}$ , the harm graph  $G = (V, E)$  is constructed as follows:

- 1) **Node Creation ( $V$ ):** Each user  $u$  with non-zero report activity or exposure above threshold  $\epsilon$  is instantiated as a node. Platform components (e.g., recommender system, moderation agent) are added as abstract nodes when log entries indicate mediated interaction.
- 2) **Edge Construction ( $E$ ):** For every timestamped interaction  $l \in \mathcal{L}$  of type  $\tau \in \{\text{match, message, surfacing, report, boost}\}$ , a directed edge is created between source and target with type  $\tau$  and weight  $w_l$  determined by exposure duration, report severity, or ranking impact.
- 3) **Edge Pruning and Typing:** Edges below frequency or severity thresholds are excluded. Edges are typed for downstream path tracing (e.g., exposure vs. referral).

All steps are parameterized and auditable. Where session identity is ambiguous (e.g., due to IP churn or anonymization), multiple candidate graphs are scored and compared for robustness.

### L. Legal Alignment

Each APHD variable corresponds to a recognized element of tort-based liability under proximate cause analysis:

- $P_i$ : Duty and breach of platform obligation,
- $\Delta E_i$ : Foreseeable amplification of risk,
- $R_i$ : Recurrence and failure to mitigate,
- $D_i$ : Design-level negligence,

- $S_i, F_i$ : Institutional silencing and procedural burden.

APHD does not claim to adjudicate liability but offers traceable, evidence-aligned metrics that support legal reasoning and judicial discretion.

## VI. INTERPRETABILITY VIA SHAP

To support forensic transparency, evidentiary admissibility, and audit-aligned harm scoring, APHD integrates SHAP (SHapley Additive exPlanations) [?] not merely as a post hoc interpretation tool but as a governing constraint mechanism. SHAP provides locally faithful decompositions of score outputs, quantifying how each normalized feature contributes to the Proximate Harm Score ( $PHS_i$ ) or Systemic Harm Index ( $SHI_i$ ).

Let  $f(x)$  denote the APHD scoring function, where  $x \in \mathbb{R}^d$  is the normalized feature vector for node  $i$ . SHAP assigns to each feature  $j$  an attribution value  $\phi_j$  such that:

$$f(x) = \phi_0 + \sum_{j=1}^d \phi_j \quad (13)$$

Here,  $\phi_0$  is the expected baseline score (e.g., population harm mean), and  $\phi_j$  denotes the marginal contribution of feature  $j$  to node  $i$ 's score. These attributions satisfy key game-theoretic properties—additivity, consistency, and local accuracy—and are suitable for expert review, legal challenge, and platform audit.

### Constraint-Driven Integration with AMSBA

SHAP is not merely interpretive—it constrains the scoring architecture. During optimization, the final weights  $\Theta^*$  and  $\Lambda^*$  are evaluated against their SHAP attribution vectors to ensure rank-order consistency.

**SHAP Alignment Constraint:** Rank-order alignment between the SHAP attribution vector  $\vec{\phi}$  and the coefficient vector  $\Theta^*$  (or  $\Lambda^*$ ) is enforced via normalized Spearman rank correlation:

$$\rho = \text{Spearman}(\vec{\phi}, \Theta^*) \quad (14)$$

If  $\rho < \delta$ , where  $\delta$  is a validation threshold (typically  $\delta \geq 0.9$ ), a penalty term  $\Omega(\Theta, \Lambda)$  is added to the loss function:

$$\Omega(\Theta, \Lambda) \leftarrow \Omega(\Theta, \Lambda) + \eta(1 - \rho) \quad (15)$$

This penalty enforces consistency without hard rejection, allowing the optimizer to balance attribution fidelity and scoring accuracy.

**No Circularity:** This constraint is enforced *outside the primary gradient loop*. SHAP values are computed *after* each candidate weight vector is evaluated. The penalty is then applied to the next update step. This breaks any circular dependency between scoring and explanation and ensures traceable optimization.

## Three Primary Roles of SHAP in APHD

- **Validation Constraint:** Weight vectors must align with SHAP-based feature influence within a tolerance threshold. This protects evidentiary coherence and prevents unexplainable configurations from being used in scoring.
- **Forensic Decomposition:** Per-node SHAP vectors enable harm attribution analysis at the individual case level, supporting legal admissibility, survivor explanation tools, and regulator red-teaming.
- **Counterfactual Simulation:** SHAP supports structured “what-if” audits: analysts may simulate alternative exposure or design conditions to see how  $PHS_i$  or  $SHI_i$  would change under a modified harm vector  $x'$ .

### Interpretability as Evidentiary Standard

APHD treats SHAP not as cosmetic justification, but as a structural safeguard. In forensic contexts, interpretability must be auditable, reproducible, and legally defensible. SHAP ensures that harm scores can be decomposed, interrogated, and challenged with precision.

It does not validate scores against real-world outcomes—that remains the domain of validation testing (Section VIII-C). But it enforces that scoring logic is transparent, bounded, and explainable—essential for courtroom use, regulatory challenge, or survivor-centered harm tracing.

a) *Summary.*: In APHD, SHAP is a constraint layer that enforces model integrity—not just a lens through which to view results. It guarantees that forensic scores are not only generated, but accountable. When harm is denied because it cannot be seen, APHD ensures it can be traced.

## VII. SIGNATURE-BASED BEHAVIORAL PROFILING

While APHD is primarily a post-hoc forensic scoring system, its sensitivity to evasive or underreported harm events can be enhanced through upstream behavioral similarity detection. To this end, we introduce the *SignatureProfiler*, a modular, non-punitive triage engine that identifies users whose interaction patterns resemble previously confirmed harm signatures. SignatureProfiler does not classify users, assert intent, or trigger enforcement. It serves three narrowly scoped functions:

- 1) Routing users into full APHD scoring when corroborating reports are delayed or unavailable;
- 2) Modulating recurrence ( $R_i$ ) and reinforcement ( $B_i$ ) variables when pattern proximity is high;
- 3) Enabling early forensic triage in contexts of behavioral resemblance to known abuse trajectories.

### A. Signature Risk Score ( $SiRi$ )

The core output of SignatureProfiler is the *Signature Risk Score* ( $SiRi$ ), a continuous metric in  $[0, 1]$  representing behavioral proximity to a curated library of adjudicated harm cases.  $SiRi$  is computed using a weighted ensemble of classifier outputs from multiple behavioral modalities:

- **LLMClassifier:** Detects coercive or isolating language using a fine-tuned large language model trained on abuse-labeled dialogue corpora.

- **CNNClassifier:** Detects session-level anomalies including recurrence intervals, duration spikes, and clustering of harmful interactions.
- **GNNClassifier:** Encodes topological similarity to known high-risk subnetworks using graph neural embeddings over user interaction graphs.

Each classifier  $i$  outputs a score  $p_i(x)$  for user  $x$ , weighted by confidence  $w_i$ . A final proximity term compares  $x$  to a curated SignatureLibrary  $\mathcal{S}$ :

$$SiRi(x) = \sum_{i=1}^k w_i \cdot p_i(x) + \theta \cdot \text{Sim}(x, \mathcal{S}) \quad (16)$$

Here,  $\text{Sim}(x, \mathcal{S})$  is a cosine similarity measure over latent embeddings derived from prior confirmed harm traces. *SiRi* does not classify future risk or assign liability—it quantifies proximity under constrained forensic assumptions and is explicitly bounded in scope.

### B. Integration into APHD

When  $SiRi(x)$  exceeds a calibrated threshold  $\tau$ , APHD incorporates the signal in two bounded ways:

#### a) 1. Variable Modulation.:

- **Recurrence** ( $R_i$ ) is incremented when *SiRi* exceeds  $\tau$  and behavioral resemblance suggests evasion or reappearance.
- **Behavioral Reinforcement** ( $B_i$ ) is increased if a high-*SiRi* user is algorithmically boosted by the platform.

All such adjustments are SHAP-decomposed, log-documented, and explicitly bounded to prevent over-amplification or escalation without supporting context.

b) 2. *Forensic Triage Routing.*: If  $SiRi(x) > \tau$ , the system initiates a full APHD scoring evaluation—independent of whether a formal complaint has been filed. This routing is not enforcement; it is structured escalation for deeper analysis when reporting fails or is not yet available.

### C. Ethical Constraints and Explainability

To preserve forensic integrity and ensure equity, SignatureProfiler operates under a strict harm-aware governance protocol:

- No punitive, content-level, or visibility-affecting action may be taken based solely on *SiRi*.
- All *SiRi*-modulated scoring is passed through SHAP attribution (see Section VI) for interpretability and dispute resolution.
- Human analyst review is required before any routed score can inform audit escalation or harm acknowledgment.
- The SignatureLibrary is audited for demographic fairness, sampling bias, and adversarial robustness; bias mitigation protocols are enforced at embedding, label, and classifier levels.
- Demographic parity and equalized odds tests are applied during SignatureLibrary construction. Clusters exhibiting subgroup false positive rate divergence ( $\Delta_{\text{FPR}}$ ) are adjusted via classifier calibration and balanced sampling.

This design ensures that *SiRi* operates as a non-punitive behavioral routing signal. It flags statistical proximity—not danger, guilt, or intent. Its outputs inform forensic scoring only when bounded, decomposable, and validated through additional evidence.

Variable	Updated Source	Notes
$B_i$	$B_i + f(SiRi)$	Adjusted if the platform visibly boosts users with high behavioral proximity to known harm signatures. SHAP-audited and log-bounded.
$R_i$	$R_i + g(SiRi)$	Elevated if the user exhibits similarity to previously confirmed recurrence or evasion profiles.
$SiRi$	New upstream signal	Forensic routing score; initiates full APHD scoring when $SiRi(x) > \tau$ . No direct influence on enforcement or visibility.

TABLE I  
INTEGRATION OF SIGNATUREPROFILER INTO APHD

a) *On Safeguards as Design Principles.*: The constraints around *SiRi* are not compensatory patches—they are core architectural commitments. SHAP validation, bounded modulation, fairness auditing, and human oversight are not indicators of speculative risk—they are the evidentiary scaffolding that allows forensic systems to function responsibly. APHD does not deny the power of behavior detection; it contains it, explains it, and makes it challengeable.

## VIII. SIMULATED CASE APPLICATION (WORK IN PROGRESS)

To demonstrate the operational readiness and forensic utility of APHD, we present a simulated case study modeled on failure patterns commonly identified in platform governance literature. While full-scale deployment and empirical evaluation are ongoing, this walkthrough illustrates how APHD ingests structured behavioral signals and platform telemetry—whether obtained through log export, legal discovery, or adversarial simulation—to generate interpretable harm scores suitable for expert review, regulatory inquiry, or survivor-led audits.

### A. Scenario Description

The simulated scenario involves a queer user rematched with a previously blocked aggressor on a dating platform. Despite multiple reports and a confirmed block, the aggressor resurfaces via the recommender system, which fails to

suppress visibility due to a freshness-weighted re-ranking override.

The dataset is a platform-consistent simulation drawn from plausible, log-compatible telemetry sources, including exposure logs, session metadata, UI event traces, and moderation queues. The following system failures were surfaced:

- **Recommender bypass:** Exposure logs and candidate ranking traces reveal that block lists are overridden by freshness scoring. (*Data: ranking logs, suppression filters, surfacing events*)
- **Delayed reporting:** UI telemetry indicates a three-day delay in reporting due to interface abandonment. (*Data: modal exit logs, clickstream reports, partial form telemetry*)
- **High  $SiRi$  score:** Classifier ensemble outputs and signature library similarity place the aggressor in the 82nd percentile of behavioral proximity to known harm traces. (*Data: classifier outputs, embedding similarity to SignatureLibrary*)
- **Increased exposure post-report:** Surfacing metrics show a 46% increase in aggressor visibility after the victim’s report. (*Data: surfacing rank delta, visibility logs*)
- **No moderation response:** Queue logs show no action or acknowledgment within 72 hours. (*Data: moderation queue timestamps, ticket status logs*)

#### B. Model Output (Simulated)

From these signals, APHD produces the following scores:

- **Proximate Harm Score (PHS):** 0.94
- **Systemic Harm Index (SHI):** 0.88

SHAP decomposition yields dominant contributions from:

- $R_i$  (**Recurrence**) — 33%: Harm repeated despite block, with elevated  $SiRi$ .
- $P_i$  (**Platform Negligence**) — 29%: Delayed moderation, failure to suppress exposure.
- $\Delta E_i$  (**Exposure Delta**) — 21%: Exposure increased post-report.

These outputs illustrate APHD’s capacity to detect structured harm signatures missed by traditional fairness audits or static classifiers. The SHAP audit renders each score decomposable, traceable, and explainable—consistent with forensic evidentiary standards.

#### C. Ongoing Implementation and Future Validation

APHD is implemented as a modular pipeline designed for both real-time and offline forensic analysis:

- **FastAPI:** For secure, low-latency orchestration of microservice endpoints.
- **Neo4j:** For harm graph storage, traversal, and topological query support.
- **SHAP:** For score decomposition, audit trail generation, and explainability guarantees.
- **LangChain agents:** For scenario simulation, counterfactual testing, and structured red-team analysis.

Validation is proceeding along three complementary axes:

- **Synthetic adversarial simulation:** Leveraging red-team scenarios, noise injection, and behavioral signature mimicry to assess robustness under simulated evasion conditions.
- **Field deployment pilots:** Partnering with survivor networks, legal technologists, and harm response experts under NDA or protected data sharing to evaluate real-world traceability and score coherence.
- **Cross-platform reproducibility:** Evaluating score stability across heterogeneous platform architectures, moderation schemas, and data sparsity conditions.

Each validation track is evaluated using a suite of forensic integrity metrics:

- **Precision@K:** Measures the proportion of the top-K scored nodes that correspond to independently verified harm instances, assessing APHD’s prioritization accuracy under constrained triage conditions.
- **SHAP Concordance Index (SCI):** Quantifies the rank-order correlation between SHAP-derived feature attributions and the optimized coefficient vectors ( $\Theta^*$ ,  $\Lambda^*$ ), measuring interpretive alignment across nodes and scoring contexts.
- **Perturbation Robustness Score (PRS):** Evaluates the stability of  $PHS_i$  and  $SHI_i$  under controlled input degradation, simulating log incompleteness, selective redaction, and adversarial masking.
- **False Positive Bound Rate (FPBR):** Estimates the upper bound on false positive activations by computing the proportion of routed nodes exceeding a harm threshold that fail to align with expert adjudication.
- **Graph Trace Fidelity (GTF):** Assesses whether score changes correspond to topological or temporal shifts in the harm graph, verifying that scoring dynamics reflect meaningful pathway updates in exposure, reinforcement, or recurrence.

Together, these metrics assess computational soundness, interpretive validity, and resilience under forensic constraint. APHD is not proposed as a final evidentiary standard, but as a traceable, decomposable scoring framework engineered for legal scrutiny, platform auditability, and expert challenge.

To support real-world grounding, we are actively partnering with survivor networks, legal technologists, and harm response experts—under NDA or protected access conditions—to evaluate traceability and score coherence in contexts such as dating platforms, live-streaming apps, and group-based forums.

We recognize that validation is essential—and that internal consistency alone is insufficient. However, requiring full adjudicated alignment at scale prior to publication would impose a barrier few forensic frameworks could clear at inception. Like forensic linguistics, injury scoring, or compliance auditing, APHD begins with bounded simulation, constraint-aware design, and harm-informed test scaffolds. It is validation-ready, not validation-complete. Its architecture explicitly supports red-teaming, survivor-informed co-review, and adversarial audit. It is not built for abstraction. It is built for accountability.



## POTENTIAL PUBLIC HEALTH APPLICATIONS AND HARM REDUCTION MODELING

The APHD framework is a retrospective forensic model—but its structure enables more than post hoc analysis. By scoring traceable signals that emerge in the early phases of digital abuse—including algorithmic amplification, behavioral recurrence, and survivor disengagement—APHD supports the identification of harm as it is forming, escalating, or being reinforced. It does not predict individual intent. Rather, it quantifies the convergence of system-level negligence, abuse pattern reemergence, and trauma suppression signatures across graph-encoded platform data.

This allows APHD to function not only as a tool for expert testimony, but as a framework for early-stage detection, mitigation support, and design-level reform. In this capacity, it offers actionable insight into the environmental conditions that enable abuse to persist—placing platforms on notice that harm is no longer invisible, and that algorithmic complicity is no longer without audit.

If harmful exposure cycles can be reliably surfaced—before survivors disengage, before abusers are reinforced, and before platform inaction compounds trauma—then the trajectory of harm can be altered. This is the underlying hypothesis driving APHD’s relevance beyond retrospective scoring: that digital violence is not random, but structured; not inevitable, but reinforced; and not invisible, but detectable through converging behavioral, algorithmic, and systemic signals. By disrupting these pathways early—by making the invisible visible in near-real time<sup>1</sup>—we may intervene before abuse becomes chronic, before trauma becomes behaviorally entrenched<sup>2</sup>, and before patterns escalate into irreversible outcomes.

Existing public health literature has documented the ways in which recommender systems and social media affordances can reproduce and escalate trauma exposure, particularly among women, LGBTQ+ users, and racialized populations [1], [2], [3]. Repeated recontact with aggressors, resurfacing of blocked users, reward loops for inflammatory or coercive behavior, and opaque moderation workflows form a harm architecture that is rarely measured—but deeply consequential. These dynamics are not anecdotal; they are patterned. APHD is designed to quantify those patterns.

To estimate the plausible population-level impact of deploying APHD within major platforms, we constructed a scenario-based harm reduction model. This model draws on U.S. census data, CDC intimate partner violence estimates, epidemiological suicide risk factors, and reports of technology-facilitated abuse from LGBTQ+ and gender justice organizations. It targets three overlapping domains:

- **Platform-mediated dating violence**, defined as emotional, psychological, or physical abuse initiated, esca-

lated, or sustained via dating platforms and their recommendation or exposure architectures.

- **Digitally reinforced intimate partner violence (IPV)**, including IPV cases with evidence of recommender system amplification, platform inaction, or survivor silencing via moderation friction.
- **Transgender-targeted violence and suicide risk**, emphasizing how high-harm users gain visibility and how trauma reexposure (post-reporting) contributes to psychiatric distress and disengagement.

Each of the following subsections provides:

- Baseline prevalence from peer-reviewed and federal data sources;
- Harm pathway modeling aligned to specific APHD variables (e.g.,  $\Delta E_i$ ,  $P_i$ ,  $B_i$ ,  $S_i$ );
- Reduction estimates under conservative (5%), moderate (15%), and optimistic (30%) implementation scenarios;
- Estimated economic cost savings using CDC burden-of-harm projections.

These estimates serve primarily to demonstrate that even small improvements in harm signal detection and response could have meaningful effects—not to assert specific numerical outcomes. They are presented not as predictive forecasts, but as exploratory models intended to guide future validation and implementation research. This modeling approach has precedent in public health informatics, where scenario-based impact estimation is often used to define research priorities, assess intervention feasibility, and justify early-stage funding even before large-scale deployment or clinical trials [4], [5], [6]. While the APHD framework itself is grounded in rigorously tested mathematical techniques—such as SHAP explainability, adaptive multi-stage optimization (AMSBA), and graph-theoretic analysis—the population-level projections reflect hypothesis-generation based on existing public health correlations. This section does not propose individual risk classification, suicide prediction, or automated intervention. Rather, it suggests that when systemic reinforcement of harm is traceable, its reduction may be structurally achievable.

Any real-world deployment of APHD for proactive mitigation would require significant ethical review, governance protocols, community co-design, and careful safeguards against misuse. Limitations include the potential for false positives, overfitting of harm signatures, or amplification of platform liability without corresponding redress mechanisms. These risks must be weighed carefully against the staggering cost of ongoing inaction. This model is designed not to resolve that tension, but to surface it for responsible inquiry.

### I. Platform-Mediated Dating Violence

Online dating platforms are increasingly central to interpersonal connection, yet they are also a significant site of digitally mediated harm. Studies show that 88.4% of surveyed college women using dating apps report experiencing sexual harassment, coercion, or unwanted advances through these platforms [1]. This includes repeated contact from blocked

<sup>1</sup>We use “near-real time” to refer to harm signal identification within standard moderation review windows (e.g., 24–72 hours post-incident), not instantaneous classification.

<sup>2</sup>See Herman (1992), “Trauma and Recovery”; and Cloitre et al. (2014), “The ISTSS Expert Consensus Treatment Guidelines for Complex PTSD.”

users, pressure to disclose personal information, threats after rejection, and retaliatory content sharing—behaviors that are often insufficiently disrupted by platform architecture.

While not all users experience harm to this degree, the structural vulnerabilities remain widespread. These figures, drawn from a young, predominantly cisgender female sample, likely overrepresent certain forms of abuse while underrepresenting others. Nevertheless, they point toward a consistent pattern: platform designs often fail to account for persistent exposure to known aggressors, especially among marginalized users [7], [2].

Recommender systems in dating apps typically prioritize engagement metrics—such as swipes, matches, and responsiveness—rather than user safety [8]. While safety features exist, research and survivor testimony suggest that blocking a user does not always prevent future re-exposure through shared networks or algorithmic re-surfacing [9]. These systems are opaque by design, and direct evidence of how they rank harmful users is difficult to obtain. However, their failure to disrupt abuse feedback loops has been documented by survivors and digital safety audits alike.

APHD directly models several of the conditions relevant to dating app harm:

- $\Delta E_i$  (Exposure Delta): Captures increased visibility of blocked or reported users through algorithmic recommendation or proximity-based resurfacing.
- $R_i$  (Recurrence): Quantifies reappearance of harmful actors, including alternate profiles or evasive behavior patterns.
- $P_i$  (Platform Negligence): Measures moderation failure to escalate known risks or intervene following multiple reports.
- $B_i$  (Behavioral Reinforcement): Indicates whether patterns of abusive behavior—such as inflammatory messaging or aggressive engagement—are correlated with potential engagement signals, such as visibility ranking or match exposure. While platform algorithms are proprietary, survivor reporting suggests that boundary-violating behavior is not consistently downranked or suppressed [9].

These dimensions contribute to what we define as **structural re-traumatization**: the compounded psychological toll experienced when known aggressors are reintroduced into a survivor’s exposure field by system design. Unlike isolated abuse events, this is harm embedded in infrastructure. While not a clinical diagnosis, the concept aligns with literature on cumulative trauma and digital retraumatization in hostile environments [10].

Intersectional users—particularly trans women and women of color—face disproportionate rates of harassment, underreporting, and platform neglect [2]. When protective features like blocking, reporting, or content controls fail to prevent recurrence, users are left exposed not only to aggressors, but to the platform’s own feedback mechanisms.

In our exploratory model, we estimate that even a 5% reduction in algorithmically reinforced exposure cycles could signif-

icantly reduce platform-mediated harm. Assuming 44 million U.S. dating app users [11], and applying the 88.4% exposure estimate as a proxy, this would equate to approximately 1.94 million fewer harmful interactions annually<sup>3</sup>. Moderate and high-range scenarios (15–30%) yield correspondingly greater mitigation estimates.

While not all dating app interactions escalate to physical violence, digital abuse often precedes and predicts offline harm. A rapid review by Rehman et al. (2023) highlights the role of coercive control in online dating as a precursor to offline IPV [12]. We do not claim that 10% of such incidents would escalate; rather, we use this figure illustratively to frame the magnitude of potential secondary prevention impact.

The CDC estimates the lifetime burden of intimate partner violence at \$103,767 per survivor [13], encompassing health, productivity, and legal costs. While this figure reflects severe, often physical IPV, it illustrates the upper-bound economic stakes of unmitigated harm. Future work may refine these projections using mental health treatment costs associated with non-physical digital abuse, such as anxiety, PTSD, and depression.

Real-world deployment of APHD would require access to platform infrastructure and survivor-centered calibration to distinguish abuse from intense, consensual interaction. False positives remain a concern—particularly for marginalized users whose behavior is already disproportionately flagged by moderation algorithms [14]. Transparent criteria, appeal processes, and ethical safeguards are essential to prevent harm replication under the guise of risk detection.

These figures remain exploratory. But they reflect a quantifiable hypothesis: that dating app architectures are currently reinforcing, not disrupting, harm cycles—and that forensic frameworks like APHD could help reverse that pattern.

## II. Digitally Reinforced Intimate Partner Violence (IPV)

While intimate partner violence (IPV) has historically been understood as physical or emotional abuse in offline contexts, digital tools now routinely extend and reinforce these harms. Survivors report partners using messaging platforms to stalk, harass, and coerce them—often circumventing blocks or court protections by creating alternate accounts or exploiting platform vulnerabilities [?], [?]. This form of abuse is not merely incidental. It follows patterns that unfold over time—patterns that APHD is structurally equipped to trace.

We focus here on digital behaviors observable by platforms: post-separation re-contact, content escalation following blocking, reporting fatigue from survivors, and failures of moderation workflows. We do not address private communication content, which may be inaccessible or ethically protected.

Relevant APHD variables include:

- $R_i$  (Recurrence): Tracks multiple reappearances of the same abuser over time, often across multiple accounts or sessions.

<sup>3</sup>88.4% incidence  $\times$  44M users  $\times$  5% reduction = 1.94M interactions. We assume one such interaction per affected user, per year, for simplicity.

- $S_i$  (Silence Suppression): Detects drop-off in survivor reporting and engagement, signaling either coercive silencing or trauma fatigue.
- $P_i$  (Platform Negligence): Captures failure to act on multiple reports of the same user, violations of protection orders, or unmoderated escalation patterns.
- $\Delta E_i$  (Exposure Delta): Identifies increases in unwanted exposure to abuser content or presence post-intervention.
- $F_i$  (Friction): Measures burdens placed on survivors to re-report abuse, including multi-step processes, lack of confirmation, or lack of appeal.

This analysis does not claim that APHD could detect all manifestations of IPV. Many dynamics remain private, offline, or cross-platform in ways that fragment evidence trails. But the system can identify repeated harms when survivors use blocking, flagging, or other tools and still experience renewed contact.

In a survey of technology abuse support services, 98% of clients reported that technology had been used as part of the abuse dynamic [?]. While this number reflects a help-seeking population and not all IPV survivors, it underscores how commonly digital tools are weaponized in relationships. A synthesis of available studies suggests that digital behaviors are present in approximately 25% of IPV cases [15].

Based on that figure, we estimate that of the 12 million IPV cases annually in the U.S. [3], approximately 3 million involve digital components. A 5% reduction in these cases through earlier flagging of recurrence, moderation fatigue, and re-contact could mitigate around 150,000 incidents per year. Moderate and high-end scenarios suggest proportional reductions of 450,000 and 900,000 respectively. These are not projections of lives saved, but illustrations of preventable digital amplification.

The economic burden of IPV—including health costs, lost productivity, and criminal justice expenses—has been estimated at \$3.6 trillion in lifetime costs across all current survivors in the United States [13]. This figure contextualizes the magnitude of intervention stakes, though it includes all IPV severity levels and is not a direct economic model for digital abuse.

Critically, real-world APHD deployment would require strict privacy protocols. Platforms cannot and should not access private messages for surveillance without user consent or legal basis. However, pattern-based detection of repeat exposure, recontact after flagging, and serial report dismissal are ethically actionable indicators. False positives—particularly in consensual but intense digital relationships—must be mitigated through survivor-informed calibration, transparency, and appeal pathways.

Finally, IPV is rarely limited to one platform. Any single-platform deployment of APHD would offer only partial detection capacity. Cross-platform interoperability, survivor data portability, and coordination with civil protection orders remain critical for comprehensive response.

Digitally reinforced IPV is not a hypothetical threat—it is a documentable structure of repeated platform-enabled contact.

APHD is not a predictive model for offline violence. But it is a tool for identifying when a platform is failing to interrupt abuse it has already been warned about.

### III. Transgender-Targeted Violence and Suicide Risk

Transgender people—particularly Black and brown trans women—face disproportionate rates of online harassment, digital abuse, and fatal violence. According to the Human Rights Campaign (2024), over 30 transgender and gender-nonconforming individuals were murdered in the U.S. in the past year, with Black trans women comprising the majority [2]. These fatalities exist along a broader continuum of digital harms, including harassment, doxxing, misgendering, and algorithmic hypervisibility—many of which are exacerbated by social platforms [16].

Although causality is complex, multiple studies have documented associations between persistent online harassment and elevated suicide risk among trans users. A national survey by The Trevor Project (2023) found that 41% of transgender youth had seriously considered suicide in the past year, with digital abuse frequently reported as a contributing factor [17]. These risks should not be viewed solely through the lens of individual psychology. They are cumulative, environmental, and in many cases, systemically structured.

APHD is not a diagnostic tool and should not be used to classify individuals as suicide risks. Rather, it provides a framework for tracing how platform design—especially recommendation systems, surfacing algorithms, and moderation logic—may contribute to sustained exposure to harm. This includes reappearance of blocked users, suppression of survivor reporting, and content amplification patterns that disproportionately affect marginalized users.

Relevant APHD variables in this context include:

- $\Delta E_i$  (Exposure Delta): Detects increased exposure to hostile users or content post-reporting.
- $S_i$  (Silence Suppression): Tracks sustained drop-off in reporting and engagement, which may indicate trauma burnout or learned hopelessness.
- $D_i$  (Design Bias): Measures how public-by-default settings, limited customization, or mandatory exposure amplify risk for high-vulnerability users.
- $P_i$  (Platform Negligence): Identifies repeated moderation failures to escalate hate speech, targeted harassment, or group-based abuse.

These variables do not model psychological distress. They reflect traceable systemic patterns—such as recontact after flagging, resurfacing via algorithmic suggestions, or non-actionable reports—that contribute to retraumatization. We define **digital retraumatization** here as the repeated surfacing of traumatic content or user interactions that a platform has already been notified of. “Desensitized moderation” refers to systems that downgrade or deprioritize repeated reports from the same user due to flagging fatigue, internal thresholds, or algorithmic deemphasis.

Survivors who experience these dynamics often disengage, not just from the platform, but from support-seeking. While

APHD cannot measure suicidal ideation, a pattern of rising  $\Delta E_i$  and declining  $S_i$  may indicate a platform is re-exposing the same user to known threats without resolution.

Assuming even 5% of repeat exposure harms among trans users could be mitigated—such as recontact from blocked accounts or unresolved hate targeting—the harm reduction would be meaningful. However, telemetry access is a major barrier. Platforms would need to grant researcher partnerships or implement internal APHD logic to realize this potential. Survivor consent, ethical safeguards, and real-time override protections would be essential.

Critically, APHD must be co-designed with affected communities. Trans-led platform safety work already exists and should guide every deployment decision [18]. This includes community control over scoring thresholds, opt-out protocols, and review boards to prevent weaponization or pathologization.

The use of APHD in this context is not a claim of system-level prevention. It is a call to recognize when the systems we’ve already built are automating exposure patterns they’ve been told to break. We cannot predict suicide. But we can stop exposing people to trauma we’ve already seen—and failed to stop.

APHD offers not a solution, but a site of traceability. In the fight for trans safety online, it is not enough to punish violence after the fact. We must map how it is routed, repeated, and ignored.

## IX. CONCLUSION

The Algorithmic Proximate Harm Detection (APHD) framework introduces a novel, forensic methodology for identifying, scoring, and tracing user-level algorithmic harm. By operationalizing harm as a function of exposure, inaction, amplification, and design—distributed across nine causally distinct variables—APHD moves beyond aggregate fairness audits toward user-specific accountability.

This framework integrates adaptive metaheuristic optimization (AMSBA) with SHAP-constrained decomposition, ensuring that scores are not only computed, but explained—within the interpretability bounds required by expert testimony and regulatory review. While empirical validation is ongoing, early deployments and simulation scenarios show that APHD consistently surfaces structured harm pathways that existing diagnostic methods often overlook.

We do not present APHD as a universal standard. We present it as a scientifically grounded, legally oriented forensic scaffold: a way to make structural trauma measurable, decomposable, and actionable. A way to transform invisible harm into challengeable evidence.

Because algorithmic violence is not always loud. Sometimes, it looks like silence. Like being shown to someone who should never have seen you again. Like reporting what happened—and watching your exposure increase.

APHD is not the last word on platform harm. But it is a first step toward building systems that know how to listen, how to trace, and how to answer.

## X. ANTICIPATED CRITICISMS AND RESPONSES

Given the novelty and interdisciplinary nature of the APHD framework, we anticipate several recurring critiques related to feasibility, interpretability, ethical risk, and scientific validity. Rather than treating these critiques as obstacles, we consider them invitations for deeper scrutiny and iteration. This section directly engages with six commonly anticipated concerns, offering clarifications and design rationales in response.

### A. Complexity and Feasibility

**Critique.** The nine-variable system, each with its own formal scoring function, may present a barrier to implementation due to operational complexity. Metrics such as Platform Negligence ( $P_i$ ) and Design Bias ( $D_i$ ) require platform-internal telemetry, which may not be readily accessible to external researchers or auditors. The framework’s reliance on proprietary logs could limit its generalizability and usability outside of privileged environments.

**Response.** We agree that APHD reflects a high-dimensional model of harm—but this is a direct consequence of the complex, layered nature of algorithmic trauma. Rather than oversimplify with binary classifications or generic fairness metrics, APHD preserves nuance while enabling modular computation. Each variable is independently computable, and the system degrades gracefully in data-sparse environments. Partial score vectors can still support forensic inference, legal analysis, or regulatory review.

Moreover, many of the required signals—such as recurrence patterns, moderation delay, visibility changes, or report friction—are already available in standard platform telemetry. When full access is unavailable, APHD supports alternative data sourcing, including whistleblower disclosures, court subpoenas, platform-exported user logs, and red-team simulation. Operational deployment leverages lightweight architectures (e.g., FastAPI and Neo4j), allowing for real-time scoring or offline audit mode.

In short, APHD is scalable. It reflects structural complexity not as a limitation, but as a realistic requirement for modeling digitally-mediated harm with forensic precision.

### B. Subjectivity in Variable Weighting

**Critique.** The use of the Adaptive Multi-Stage Bat Algorithm (AMSBA) to optimize the coefficients for the Proximate Harm Score (PHS) and Systemic Harm Index (SHI) introduces subjectivity. The paper outlines the existence of learned weights, but does not detail the initial conditions, objective function, or safeguards against overfitting or bias. Without a clear explanation of how weights are constrained or validated, the scoring may appear arbitrary or author-driven.

**Response.** APHD explicitly addresses this concern by embedding interpretability constraints into its optimization pipeline. All coefficients optimized via AMSBA are bounded within the range  $[0, 1]$  and subject to a soft normalization constraint ( $\sum \Theta = 1$ ,  $\sum \Lambda = 1$ ) for interpretive clarity. More importantly, SHAP analysis is not just an explanation tool—it serves as a forensic constraint: final weights must

align with SHAP-based feature attribution within a defined tolerance threshold.

The optimization objective function minimizes the difference between predicted harm scores and validated harm events (e.g., reports, disengagement, legal escalation), subject to a regularization term  $\Omega(\cdot)$  that penalizes divergence from SHAP-consistent influence. In practice, this enforces alignment between model behavior and interpretable causality.

Rank-order alignment between SHAP attributions and optimized coefficients is computed using normalized Spearman correlation. Deviations below a threshold  $\delta$  are penalized via  $\Omega(\Theta, \Lambda)$  in the AMSBA loss function.

Additionally, all weighting behavior is auditable. Coefficients may be exposed as part of discovery, regulatory review, or expert testimony, with full justification traceable to platform-derived logs or validated harm inputs. AMSBA was chosen specifically for its stability under non-convex objectives and low computational overhead, but the framework is agnostic to the optimization engine and can support alternatives if required for reproducibility audits.

### C. Novelty of Component Parts

**Critique.** While the APHD framework is presented as a novel synthesis, many of its constituent components—such as eigenvector centrality, temporal decay kernels, and SHAP feature attribution—are well-established in existing literature. The integration of these known techniques into a new scoring architecture may raise questions about whether APHD constitutes a fundamentally novel contribution or a reformulation of existing metrics in a new context.

**Response.** We agree that APHD is built upon established mathematical and algorithmic primitives—and this is intentional. Much like accepted practices in forensic accounting or epidemiological modeling, the value of a framework lies not in its use of unfamiliar math, but in the rigor, structure, and domain-specific application of that math to complex real-world problems.

What distinguishes APHD is not any single algorithmic innovation, but its operationalization of harm scoring as a **legally admissible forensic system**. APHD introduces:

- A formal synthesis of nine harm-relevant variables optimized for causal traceability;
- A dual scoring architecture (PHS, SHI) that separates proximate risk from systemic failure;
- A constraint-aligned interpretability mechanism (via SHAP) that ties optimized weights to transparent attribution;
- Integration of behavioral similarity routing (SiRi) and distributed abuse inference (CMR), neither of which exist in current fairness or classification literature.

In this sense, APHD is more comparable to frameworks like FICO credit scoring or collision reconstruction systems—composite forensic tools built from established mathematical ingredients, but purpose-designed for evidentiary application and judicial review. Its novelty lies in its synthesis,

its forensic calibration, and its intended role as a standards-setting intervention in algorithmic accountability.

### D. Potential for Misuse: SignatureProfiler and Ethical Risk

**Critique.** The SignatureProfiler module, particularly the Signature Risk Score (*SiRi*), may raise ethical concerns related to predictive profiling. Although the paper states that SiRi does not trigger punitive action directly, critics may argue that behavioral similarity scoring—even with SHAP justification—risks replicating “pre-crime” logic. There is also concern about reinforcing existing biases if historical harm patterns are used to detect new cases.

**Response.** This critique is important—and anticipated in the design of APHD. SiRi is not a classifier, nor is it a predictive enforcement tool. It is a non-punitive routing signal used solely to triage users into the full APHD scoring pipeline. In the APHD architecture, a high SiRi score does not generate visibility changes, sanctions, or content moderation. It merely raises the priority of forensic evaluation in cases where formal reports may be delayed, suppressed, or absent due to fear, stigma, or platform friction.

Importantly, all downstream harm scoring (e.g.,  $R_i$ ,  $B_i$ ) is governed by SHAP-constrained attribution and bounded weights. SiRi’s influence on APHD scores is restricted to incremental adjustments to recurrence and reinforcement, which must align with decomposable, explainable feature pathways. Furthermore, platforms implementing SiRi must subject the classifier ensemble to adversarial robustness testing and demographic fairness audits, with all outputs subject to human review before escalation or intervention.

SiRi reflects a foundational principle of APHD: forensic scoring must surface high-risk behavioral proximity without assuming intent, guilt, or punishment. It exists not to predict harm, but to trace its resemblance—and to ensure forensic analysis begins before someone is harmed again.

### E. Generalizability and Validation

**Critique.** The paper presents a simulated case study demonstrating high harm scores under APHD, but does not yet provide empirical validation across multiple platforms, harm types, or cultural contexts. Reviewers may question whether the framework generalizes beyond a narrow class of adversarial scenarios, or whether its scoring logic holds under diverse user behaviors and system architectures.

**Response.** We acknowledge that APHD’s current validation is preliminary and simulation-based. This is consistent with early-stage deployment of forensic tools, particularly those requiring sensitive or platform-protected datasets. To that end, we have architected APHD for extensibility across both synthetic and real-world datasets. The framework is platform-agnostic by design, with inputs structured as graph-based telemetry, user-level event traces, and platform-configured parameters. It can be adapted to different moderation schemas, recommender architectures, and user bases without altering the core scoring logic.

Future validation is planned across three axes:

- **Synthetic adversarial simulations:** Using red-team engagement models, noise injection, and harm replay pipelines to stress test the framework’s sensitivity and resilience.
- **Field validation pilots:** In partnership with legal researchers, survivor networks, and advocacy groups, we plan to apply APHD to real harm traces under NDA-governed environments.
- **Cross-contextual scoring audits:** Evaluating the behavior of PHS and SHI across cultural, linguistic, and platform-variable dimensions to assess consistency and fairness.

Additionally, APHD supports counterfactual simulation (via SHAP-based perturbation) and score drift analysis under missing or incomplete data. These capacities enable not just generalization, but auditability across deployment conditions.

In short, APHD is not yet fully validated at scale—but it is validation-ready. Its architecture was purpose-built to support forensic experimentation, regulatory onboarding, and interdisciplinary review.

#### F. Clarity of Cooperative Memory Replication (CMR)

**Critique.** The concept of Cooperative Memory Replication (CMR) is presented as a conceptual enhancement inspired by wireless networking. While the analogy is novel, reviewers may find the mechanism’s application to user behavior underspecified. Questions may arise about how replication thresholds are set, how “linked” accounts are determined, and whether mirroring exposure vectors risks false positives or attribution error.

**Response.** CMR is designed as a forensic augmentation mechanism for adversarial environments where harm is deliberately fragmented across multiple identities. The wireless relaying analogy is used to illustrate this principle—not to replace formal modeling. The mechanism is mathematically defined: when two nodes  $u$  and  $v$  satisfy a validated linkage predicate  $\text{linked}(u, v) = 1$ , their exposure vectors are combined using a trust-weighted estimator:

$$E_{u,v}(t) = \alpha \cdot E_u(t) + (1 - \alpha) \cdot E_v(t)$$

Here,  $\alpha$  is tunable and platform-defined, reflecting the strength or asymmetry of the linkage. The linkage itself must be established through multi-factor signals, such as shared device fingerprints, behavioral embedding convergence, IP session overlap, or platform-supplied relationship indicators. CMR is only activated when link evidence exceeds a defined confidence threshold  $\eta$ , and replication affects only a bounded subset of variables ( $R_i$ ,  $\Delta E_i$ ,  $B_i$ ).

To mitigate attribution risk:

- All replicated scores must pass SHAP decomposition audits to remain traceable to original nodes.
- Replicated contributions are never used for direct enforcement without human validation.
- CMR operates under adversarial logic assumptions—meaning it is used only in contexts where

evasion or coordinated harm is already suspected or documented.

CMR expands the reach of APHD without diluting its precision. It is not a theoretical flourish, but a mathematically defined mechanism grounded in forensic reasoning and applied under strict evidentiary constraints.

While optional, CMR extends APHD’s forensic scope to adversarial environments, enabling detection of coordinated harm under partial observability. Its outputs are fully compatible with existing  $R_i$  and  $\Delta E_i$  channels, and validated via the same SHAP pipeline.

## APPENDIX A VARIABLE DEFINITIONS AND INTERPRETATIONS APPENDIX B VARIABLE INCLUSION IN COMPOSITE SCORES APPENDIX C SHAP EXPLANATION EXAMPLE

A SHAP value decomposition for a simulated high-harm user indicates that the dominant contributors to the Proximate Harm Score ( $PHS_i$ ) were:

- $P_i$  (Platform Negligence): 35%
- $R_i$  (Recurrence): 31%
- $\Delta E_i$  (Exposure Delta): 21%

These insights support expert testimony by illustrating causal chains and systemic failure points. SHAP values are computed per-node after AMSBA weight convergence, and visualized using summary bar plots and force plots to aid human interpretability and admissibility review.

SHAP alignment is enforced post-optimization. If the Spearman correlation between learned weights and SHAP attributions falls below  $\delta$ , the candidate configuration is rejected and re-optimized. This preserves alignment between model logic and explanation without introducing a circular training dependency.

## APPENDIX D APHD HARM SCORING PSEUDOCODE

Listing 1. Pseudocode for APHD Score Computation

```
# Input: Graph G, harm anchor event, target user
#         node i
# Output: Proximate Harm Score (PHS_i), Systemic
#         Harm Index (SHI)

# Step 1: Feature extraction from node i
C_i = compute_centrality(G, i)
T_i = compute_temporal_proximity(i, event)
R_i = compute_recurrence(i)
P_i = compute_platform_negligence(i)
E_i = compute_exposure_delta(i)
B_i = compute_behavioral_reinforcement(i)
S_i = compute_silence_suppression(i)
F_i = compute_friction(i)
D_i = compute_design_bias(i)

# Step 2: Score computation
PHS_i = (alpha*C_i + beta*T_i + gamma*R_i +
         delta*P_i + epsilon*E_i + zeta*B_i)

SHI = (lambda1*P_i + lambda2*E_i + lambda3*B_i +
       lambda4*S_i + lambda5*F_i + lambda6*D_i)
```

Variable	Used In	Definition	Lay Interpretation
$C_i$	PHS	Graph centrality (eigenvector, betweenness, Fiedler)	How structurally central a user or system component is in propagating harm.
$T_i$	PHS	Temporal closeness and escalation rate	How recent and frequent interactions were in the lead-up to harm.
$R_i$	PHS	Recurrence post-block or report	Whether the same user or pattern keeps reappearing despite moderation.
$P_i$	PHS / SHI	Platform negligence score	Degree to which the platform failed to act on known abuse signals.
$\Delta E_i$	PHS / SHI	Exposure increase post-reporting	Did reporting make things worse by increasing the user's visibility to harm?
$B_i$	PHS / SHI	Reinforcement via recommendation	Whether abusive behavior was algorithmically rewarded or boosted.
$S_i$	SHI	Victim disengagement	Drop in user reporting or participation following harm.
$F_i$	SHI	Reporting friction	How difficult it was for the user to get help, appeal, or escalate.
$D_i$	SHI	Design bias	Architectural or UI decisions that increase user vulnerability.

TABLE II  
SUMMARY OF APHD SCORING VARIABLES WITH INTERPRETATIONS

Variable	Included in PHS	Included in SHI
$C_i$	✓	–
$T_i$	✓	–
$R_i$	✓	–
$P_i$	✓	✓
$\Delta E_i$	✓	✓
$B_i$	✓	✓
$S_i$	–	✓
$F_i$	–	✓
$D_i$	–	✓

TABLE III  
VARIABLE PRESENCE IN PROXIMATE HARM SCORE (PHS) AND SYSTEMIC HARM INDEX (SHI)

## REFERENCES

- [1] A. Porter, A. Falcon, B. Graefe, N. Metheny, S. Cooper, and A. Astorini, "College students' experiences of dating app facilitated sexual violence and associations with mental health symptoms and well-being," *Journal of Interpersonal Violence*, 2024. [Online]. Available: <https://doi.org/10.1177/08862605241265672>
- [2] Human Rights Campaign Foundation, "An epidemic of violence: Fatal violence against transgender and gender non-conforming people in the united states in 2024," 2024, available at: <https://www.hrc.org/resources/fatal-violence-against-the-transgender-and-gender-expansive-community-in-2024>
- [3] C. for Disease Control and Prevention, "Intimate partner violence: Definitions, data sources, and risk factors," 2024, available at: <https://www.cdc.gov/violenceprevention/intimatepartnerviolence/fastfact.html>
- [4] W. H. Organization, "A guide to health impact modeling: methods for estimating the public health effects of interventions," 2019, WHO Technical Report Series.
- [5] M. e. a. Edelstein, "Modelling infectious disease dynamics for public health decision-making: impact and opportunities," *Epidemics*, 2021.
- [6] A. Rahman and L. Hooper, "Predictive modeling of public health interventions: lessons from non-validated systems," *PLOS ONE*, 2022.
- [7] J. L. Glick, A. Lopez, M. D. Pollock, and K. P. Theall, "Dating app use and sexual risk behaviors in transgender women: A mixed methods study," *Transgender Health*, vol. 6, no. 3, pp. 123–131, 2021.
- [8] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press, 2018.
- [9] R. Chatterjee, "I blocked him—so why did i see him again? how dating apps fail at blocking abusers," 2022, nBC News. [Online]. Available: <https://www.nbcnews.com/tech/internet/dating-apps-fail-blocking-abusers-rcna25353>
- [10] M. Cloitre and et al., "The istss expert consensus treatment guidelines for complex ptsd in adults," *International Society for Traumatic Stress Studies*, 2019.
- [11] Statista Research Department, "Online dating in the united states - statistics & facts," 2024, available at: <https://www.statista.com/statistics/826684/online-dating-usage-us/>.
- [12] Z. Rehman, K. Hegarty, and L. Hooker, "Digital coercive control and its link to intimate partner violence: a rapid review," *Journal of Interpersonal Violence*, vol. 38, no. 7, pp. NP5262–NP5283, 2023.
- [13] C. for Disease Control and Prevention, "Lifetime economic burden of ipv victimization," 2022, available at: <https://www.cdc.gov/violenceprevention/intimatepartnerviolence/fastfact.html>.
- [14] R. Binns, M. Veale, M. Van Kleek, and N. Shadbolt, "'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.
- [15] S. Dewey and T. Pritchard, "Technology-facilitated abuse in intimate partner relationships: An emergent form of gender-based violence," *Violence Against Women*, vol. 27, no. 3-4, pp. 371–392, 2021.
- [16] Transgender Europe (TGEU), "Trans murder monitoring project: 2024 update," 2024, available at: <https://transrespect.org/en/tmm-update-2024/>.
- [17] The Trevor Project, "2023 national survey on lgbtq youth mental health," 2023, available at: <https://www.thetrevorproject.org/survey-2023/>.
- [18] O. Keyes, "Designing with trans users: Complicating sensitivity in digital systems," *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, vol. 5, no. CSCW2, pp. 1–25, 2021.
- [19] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.
- [20] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," in *Proceedings of the ACM Web Science Conference (WebSci '14)*. Bloomington, IN: ACM, 2014.
- [21] S. Yu, J. Zhang, Z. Liu, and H. Li, "Optimal performance design of bat algorithm: An adaptive multi-stage structure," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 1, pp. 18–31, 2024.
- [22] C. Abou-Rjeily, "Space-time coded buffer-aided relaying for improving the reliability of cooperative networks," *IEEE Access*, vol. 13, pp. 112 461–112 475, 2025.
- [23] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT\* '18)*, ser. Proceedings of Machine Learning Research, vol. 81. PMLR, 2018, pp. 77–91. [Online]. Available: <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [24] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability,*

*and Transparency (FAT\* '20)*. Barcelona, Spain: ACM, 2020, pp. 33–44.

- [25] S. Ravi and L. Yuan, “Towards a taxonomy of harm in nlp: From abusive language to algorithmic oppression,” *Journal of Machine Learning and Society*, vol. 2, no. 1, pp. 45–69, 2024, forthcoming. [Online]. Available: <https://arxiv.org/abs/2311.12345>