# Beyond Age Gates:

## A Privacy-Preserving Architecture for Predatory Pattern Detection in Digital Messaging

Travis Gilly

*Real Safety AI Foundation*

*Cary, Illinois, United States | t.gilly@ai-literacy-labs.org*

February 2026

## Abstract

Biometric age verification has become the default industry response to child safety concerns in digital communications, despite fundamental structural failures in legal compliance, technical resilience, and conceptual design. This paper presents an alternative architecture: an on-device, deterministic, rule-based behavioral pattern detection system that identifies predatory communication sequences by analyzing the sender's output patterns alone, without determining the recipient's age or identity, without collecting biometric data from any user, and without transmitting any data from the device to external infrastructure.

The system uses finite-state machines, weighted rule graphs, and deterministic input normalization to detect documented predatory behavioral sequences including grooming cadence, isolation tactics, age-probing, boundary testing, and escalation patterns. A unidirectional cryptographically signed update mechanism enables rule evolution without creating any data egress channel.

A comprehensive prior art survey across patent databases, academic literature, and commercial products confirms that no existing system combines all of these architectural properties. The architecture is presented as a platform-agnostic, embeddable software component positioned as a direct replacement for biometric age verification.

*Keywords:* child safety, predatory behavior detection, age verification alternative, on-device processing, deterministic pattern matching, privacy-preserving architecture, grooming detection, COPPA compliance

# 1. Introduction

The dominant paradigm for protecting minors in digital communications is age verification. Platforms require users to submit biometric data or government-issued identification to third-party vendors, confirming minimum age thresholds before granting access.

This paper argues that age verification is a category error. It attempts to determine who a user is as a proxy for determining whether that user will engage in harmful conduct. A verified 35-year-old who passes a facial scan receives unrestricted platform access, including interaction with minors. Age verification confirms demographics; it does nothing to identify, predict, or prevent predatory behavior.

We present an alternative that asks a fundamentally different question. Instead of "How old is this user?" the system asks "Is this user behaving like a predator?" This reframing eliminates the need for biometric collection, vendor relationships, and recipient identity determination. Detection is grounded in what the sender does, not in who the recipient is, consistent with legal doctrine holding that predatory intent is established by the offender's conduct pattern regardless of whether the target is actually a minor.

The architecture is deterministic rather than probabilistic, on-device rather than cloud-based, and sender-focused rather than demographic-dependent. It is designed as an embeddable, platform-agnostic SDK that achieves the safety objectives of age-gating mandates without the privacy and compliance costs generating opposition to biometric approaches.

# 2. Background: The Failure of Age Verification

## 2.1 Legal Compliance Failures

The Children's Online Privacy Protection Act (COPPA), as amended with updates effective June 23, 2025, explicitly defines biometric identifiers as personal information requiring verifiable parental consent before collection from children under 13 [4]. The FTC specifically declined to create an exception for biometric data collected temporarily for age verification during the 2024 rulemaking process, despite direct industry requests. Platforms serving minors by design that require facial scanning from all users are therefore collecting biometric personal information from children without requisite consent. Similar failures arise under GDPR, the UK Age Appropriate Design Code, and state statutes including the Illinois Biometric Information Privacy Act.

## 2.2 Technical Vulnerability

Age verification technologies relying on facial analysis have been demonstrated to be trivially defeatable. In documented cases from the UK Online Safety Act rollout, systems from at least two major vendors were defeated on deployment day by users pointing cameras at photorealistic video game characters. Age verification systems also transmit data to third-party vendors representing additional attack surfaces; in September 2025, one vendor experienced a breach exposing approximately 70,000 government identification documents.

### 2.3 The Conceptual Error

The fundamental problem is conceptual. Age verification answers the wrong question. The relevant safety question is not "Is this user over 18?" but "Is this user engaging in predatory behavior?" These questions are orthogonal. Most adults are not predators, and age verification clears them equally. What is needed is an architecture that detects dangerous conduct directly, without demographic proxies, biometric collection, or surveillance infrastructure.

*Figure 1: Conceptual Comparison of Safety Paradigms*

| | Age Verification Paradigm | Behavioral Detection Paradigm |
|---|---|---|
| **Question** | "How old is this user?" | "Is this user behaving like a predator?" |
| **Input** | Facial scan, government ID, or biometric template from ALL users | Sender's message patterns (text, timing, sequence) from flagged conversations only |
| **Data Flow** | User device -> third-party vendor -> platform server -> centralized database | Stays entirely on device; nothing transmitted |
| **Catches** | Minors trying to access age-restricted content | Adults exhibiting predatory behavioral patterns |
| **Misses** | Every verified adult who is actually a predator | Predators who never exhibit detectable patterns |
| **Compliance** | Triggers COPPA, GDPR, BIPA, UK AADC obligations for ALL users | No regulated data transaction occurs |

## 3. Related Work

### 3.1 Server-Side Machine Learning Approaches

The predominant approach relies on ML classifiers on server-side infrastructure. Microsoft's Project Artemis provides server-side grooming probability scoring deployed via Thorn [7]. Vogt, Leser, and Akbik formally defined the early sexual predator detection (eSPD) task using BERT-based classifiers and sliding window evaluation [8]. Commercial products (Bark, Net Nanny, Qustodio) upload content to cloud services for classification. All require message content to leave the device, depend on probabilistic classification, and typically require knowledge of both sender and recipient characteristics.

### 3.2 Client-Side Scanning Proposals

EU CSA Regulation proposals discuss client-side scanning using ML classifiers with results reported to centralized authorities [11]. CSS as proposed is on-device in processing location only, not in data containment.

### 3.3 Closest Prior Art

IBM/Activision (U.S. Patent 8,099,668 B2) implements a client-side "child monitor" with rule-based definitions in virtual environments, but reports abuse data to a server-side tracker and

assumes the protected user is already identified as a child [1]. Street and Olajide (ICCWS 2023) present on-device OCR/regex phrase matching that can operate offline, but perform keyword blocking rather than behavioral sequence detection [3]. U.S. Patent Application 2021/0234823 A1 (Anti-Toxin Technologies) models predatory behavior through progressive confidence scoring, but is entirely server-side, ML-based, and requires both sender and target age inference [2].

No identified system combines on-device containment, deterministic behavioral sequence detection, sender-only analysis, and age verification replacement.

## 4. System Architecture

The system comprises four functional components executing within the local runtime of the messaging client. The system is architecturally constrained to execute without requesting network permissions for analysis operations.

*Figure 2: System Architecture and Data Flow*

| 1. MESSAGE INTERCEPT | 2. PATTERN ANALYSIS ENGINE | 3. ACTION ENGINE | 4. UPDATE MECHANISM |
|---|---|---|---|
| Captures sender messages from chat pipeline | Input Normalization Layer (homoglyphs, leetspeak, whitespace) | Visual alerts to device user | Downloads signed rule packages |
| Passes text + local timestamps to PAE | Behavioral Sequence FSMs (grooming stages) | Conversation flagging | Verifies cryptographic signatures |
| Volatile memory only; no persistent storage | Weighted Rule Graph (fixed expert weights) | Optional message blocking | Strictly unidirectional (download only) |
| No network operations | Temporal Pattern Analyzer (cadence, velocity) | Optional P2P parent alert (category only) | Zero telemetry, zero device IDs sent |

*All components execute on-device. No data crosses the device boundary except optional parent notification (detection category only) and inbound rule updates.*

### 4.1 Message Intercept Layer

Receives outgoing and/or incoming message content from the messaging client's internal pipeline prior to display or transmission. Operates within the application's process space without network operations. Intercepted content is held in volatile memory only, never written to persistent storage.

### 4.2 Pattern Analysis Engine

The core detection component, logically constrained to accept as input only data structures associated with the sender's message content and local system time, structurally preventing recipient profile data from entering risk scoring.

#### 4.2.1 Input Normalization Layer

Applies deterministic text normalization to counter adversarial obfuscation: homoglyph substitution (mapping visually similar Unicode characters to canonical ASCII), Levenshtein distance calculation against indicator dictionaries, whitespace/punctuation reduction, and leetspeak normalization. All operations are deterministic, producing identical outputs for identical inputs without probabilistic models.

#### 4.2.2 Behavioral Sequence Finite-State Machines (BS-FSM)

A set of finite-state machines modeling documented predatory behavioral sequences. Each defines states corresponding to predatory conduct stages (initial contact, rapport building, trust establishment, isolation attempts, boundary testing, desensitization, explicit escalation, meeting solicitation) with deterministic transition rules based on normalized linguistic, temporal, and sequential patterns in the sender's output.

#### 4.2.3 Weighted Rule Graph (WRG)

A directed acyclic graph where nodes represent behavioral indicators and edges represent sequential/co-occurrence relationships with fixed weights based on forensic and criminological research. Computes composite risk scores through deterministic summation along observed behavioral paths. Weights are updateable only through the unidirectional mechanism described in Section 4.5.

#### 4.2.4 Temporal Pattern Analyzer (TPA)

Tracks sender timing characteristics including message frequency, time-of-day patterns, session duration, and escalation velocity using deterministic threshold rules derived from forensic grooming cadence literature [9].

### 4.3 Behavioral Indicator Taxonomy

The system detects the following sender behavioral indicator categories:

- Grooming Cadence: Gradually escalating intimacy, progressive disclosure solicitation, compliment escalation, emotional dependency building.
- Isolation Tactics: Attempts to move to private channels, discourage disclosure to trusted adults, establish exclusive relational framing.
- Age-Probing Sequences: Inquiries about age, school grade, physical development, especially combined with other indicators.
- Boundary Testing: Introduction of sexual/romantic/intimate content at incrementally increasing levels.
- Authority/Trust Exploitation: Positioning as mentor or confidant combined with requests for compliance or secrecy.
- Meeting Solicitation: Attempts to arrange in-person contact or migrate to real-time channels.

- Desensitization Sequences: Progressive normalization of inappropriate content across multiple interactions.

### 4.4 Action Engine

Executes configurable local responses when thresholds are met: visual alerts with configurable urgency, conversation flagging, optional message blocking, optional local-only event logging (triggers and timestamps, not content), and optional parent/guardian notification via direct encrypted peer-to-peer channel containing only detection category.

### 4.5 Unidirectional Definition Updates

Predatory tactics evolve. The system addresses this through a unidirectional rule update mechanism preserving zero-data-egress. Rule definitions are stored as updateable assets (analogous to antivirus signatures) rather than hardcoded logic. Update packages are cryptographically signed; the SDK verifies signatures before ingestion. The channel is strictly unidirectional: download only. No device identifiers, telemetry, confirmations, or user data are transmitted to the update server.

## 5. Design Principles

### 5.1 Sender-Only Detection

The system analyzes only the sender's behavioral output. No data structure representing the recipient is accessed, created, or referenced by risk scoring. This constraint reflects established legal doctrine: in the United States and United Kingdom, predatory intent is established by conduct pattern regardless of target's actual age [5, 6]. Law enforcement undercover operations demonstrate this operationally; defendants are convicted based on behavioral patterns when communicating with adult officers posing as minors.

### 5.2 Deterministic Operation

The choice of deterministic logic over ML provides reproducibility (identical inputs yield identical outputs), auditability (every decision traces to specific rules), adversarial resistance (not susceptible to adversarial examples or data poisoning), and eliminates the ethical concerns of collecting child exploitation material for training purposes.

### 5.3 Privacy by Architecture

No message content is transmitted from the device. No vendor accesses any data. No centralized database exists or can exist. No data outside volatile memory could be subpoenaed, breached, or misused. COPPA, GDPR, BIPA, and related obligations do not apply because no regulated data transaction occurs.

## 6. Comparison with Existing Approaches

*Table 1: Architectural Feature Comparison*

| System | On-Device | Determ. | Sender | No AgeV | Plat.Ag | UniUpd | InNorm | Legal |
|---|---|---|---|---|---|---|---|---|
| Proposed | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| IBM 8,099,668 | Partial | Yes | No | Partial | No | No | No | No |
| US 2021/0234823 | No | No | No | No | No | No | No | No |
| Proj. Artemis | No | No | No | No | No | No | No | No |
| Street/Olajide | Yes | Yes | No | Partial | Yes | No | No | No |
| Bark/Net Nanny | No | No | No | No | No | No | No | No |
| EU CSS Prop. | Partial | No | No | No | No | No | No | No |

No identified system satisfies all eight requirements. IBM comes closest but fails on data containment (server reporting), sender-only analysis (assumes child recipient), and platform generality (gaming-specific). Street and Olajide achieve on-device deterministic operation but perform keyword blocking, not behavioral sequence detection.

# 7. Discussion

## 7.1 Limitations

This paper presents an architecture, not a deployed product. Rule authoring requires collaboration with forensic psychologists, criminologists, and law enforcement. Deterministic systems face inherent tension between precision and recall. Platform integration requires developer cooperation. The system detects grooming-pattern predation but does not address all online harms to minors.

## 7.2 Political Positioning

A notable property is the absence of natural political opposition. Biometric systems face resistance from civil liberties organizations (privacy), children's advocates (biometric collection from minors), technologists (implementation failures), and legal scholars (compliance impossibilities). The proposed system collects no biometric data, creates no surveillance infrastructure, involves no vendors, and transmits no data. Organizations typically opposed to safety mandates and those advocating for them could endorse the same solution.

## 7.3 Implications for End-to-End Encryption

Server-side scanning is impossible in E2EE environments. Client-side scanning proposals create data egress channels undermining encryption guarantees. The proposed system operates entirely on-device with no egress, making it compatible with E2EE without compromising encryption. This may represent the only viable approach to predatory detection in encrypted messaging that does not weaken the encryption itself.

# 8. Conclusion

Age verification asks the wrong question. It asks who the user is instead of what the user is doing, collecting biometric data from children to protect them from adults who pass the same checks without difficulty.

The architecture presented here inverts that logic. By analyzing sender behavioral output through deterministic, on-device pattern matching, the system identifies the conduct that constitutes the threat without requiring any information about the recipient. No biometric data is collected. No data leaves the device. No compliance obligation is triggered.

The prior art survey confirms no existing system combines these properties. The specific combination of strict on-device containment, deterministic behavioral sequence detection, sender-only analysis, input normalization, unidirectional rule evolution, and legal doctrine alignment has not been previously disclosed as a unified architecture.

The choice between protecting children and protecting privacy is a false dilemma, an artifact of approaches that asked the wrong question from the beginning.

## Acknowledgements

This paper was produced with the assistance of speech-to-text transcription and Claude (Anthropic) as assistive technology for ADHD and ASD accommodation in organizing and structuring content. All research, system architecture, design decisions, analysis, and conclusions are solely the author's own.

## References

[1] International Business Machines Corp. (2012). Predator and abuse identification and prevention in a virtual environment. U.S. Patent 8,099,668 B2. Filed January 7, 2008.

[2] Anti-Toxin Technologies, Inc. (2021). Detecting and identifying toxic and offensive social interactions in digital communications. U.S. Patent Application 2021/0234823 A1. Published July 29, 2021.

[3] Street, J., & Olajide, F. (2023). Evaluating a non-platform-specific OCR/NLP system to detect online grooming. Proceedings of the 18th International Conference on Cyber Warfare and Security (ICCWS 2023), Towson University. Academic Conferences International Limited.

[4] Federal Trade Commission. (2024). Children's Online Privacy Protection Rule; Final Rule. 16 CFR Part 312. Federal Register, 89 FR 2034.

[5] Crown Prosecution Service. (2019). CPS strengthens guidance on child abusers caught in undercover stings. CPS Policy Guidance, United Kingdom.

[6] U.S. Department of Justice. (Various). FBI undercover operations leading to federal charges for individuals attempting to meet minors for sex. DOJ Press Releases, Southern District of Illinois and others.

[7] Microsoft/Thorn. (2020). Project Artemis: Grooming detection technology for chat platforms. Deployed via Thorn's Safer product for platform integration.

[8] Vogt, M., Leser, U., & Akbik, A. (2021). Early detection of sexual predators in chats. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and

the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), Volume 1: Long Papers, pp. 4985-4999.

[9] Bours, P., & Kulsrud, H. (2019). Detection of cyber grooming in online conversation. IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2019).

[10] Chehbouni, K., et al. (2025). Enhancing privacy in the early detection of sexual predators through federated learning and differential privacy. arXiv preprint arXiv:2501.12537.

[11] Technology Coalition. (2021). Online grooming: Considerations for detection, response, and prevention. Research Paper.