REAL SAFETY AI FOUNDATION · WORKING PAPER

CHILD SAFETY · PREDATORY BEHAVIOR DETECTION · AGE VERIFICATION ALTERNATIVE · ON-DEVICE PROCESSING · DETERMINISTIC PATTERN MATCHING · PRIVACY-PRESERVING ARCHITECTURE · GROOMING DETECTION · COPPA COMPLIANCE

Beyond Age Gates

A privacy-preserving architecture for predatory pattern detection in digital messaging.

Travis Gilly · Real Safety AI Foundation · Cary, Illinois

Working paper · February 2026

THE QUESTION AGE VERIFICATION ASKS

"How old is this user?"

Confirms demographics. A verified adult who passes a facial scan receives unrestricted access, including interaction with minors.

THE QUESTION THIS PAPER ASKS

"Is this user behaving like a predator?"

Detects conduct directly. No biometric collection, no vendor, no recipient identity determination.

Age verification answers the wrong question.

It attempts to determine who a user is as a proxy for whether that user will engage in harmful conduct. The two are orthogonal: most adults are not predators, and the check clears them all equally. Detection should be grounded in what the sender does, not who the recipient is, consistent with legal doctrine holding that predatory intent is established by conduct pattern regardless of whether the target is actually a minor.

SOURCES FTC, Children's Online Privacy Protection Rule; Final Rule, 16 CFR Part 312, 89 FR 2034 (2024) [4]. · Crown Prosecution Service Policy Guidance (2019) [5].

PART ONE

The Failure of Age Verification

Three structural failures: legal compliance, technical resilience, and conceptual design.

I

Biometric age verification became the default industry response to child-safety concerns. It fails on all three dimensions at once, and the conceptual failure is the one the others follow from.

To verify a child's age, the platform must first collect the child's biometric data.

COPPA defines biometric identifiers as personal information requiring verifiable parental consent before collection from children under 13. In its 2024 rulemaking the FTC specifically declined to create an exception for biometric data collected temporarily for age verification, despite direct industry requests. A platform requiring facial scanning from all users is therefore collecting regulated biometric data from children without the requisite consent.

THE SAME WALL, FOUR TIMES

COPPA · 16 CFR 312

Biometric identifiers = personal information; no age-verification exception.

GDPR

Biometric data is a special category under Article 9.

UK AADC

Age Appropriate Design Code; data minimization for minors.

Illinois BIPA

Written consent before biometric capture; private right of action.

SOURCES FTC, Children's Online Privacy Protection Rule; Final Rule, 16 CFR Part 312, 89 FR 2034 (2024) [4]. · Updates effective June 23, 2025.

Defeated on deployment day. Breached within months.

Day 1

During 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.

70,000

Government identification documents exposed in a September 2025 breach of a single age-verification vendor. Every third-party transmission is new attack surface.

0

Predators stopped by a check that confirms only demographics. A verified adult passes and proceeds with unrestricted access.

SOURCES Documented UK Online Safety Act rollout incidents (2025). · September 2025 age-verification vendor breach (≈70,000 ID documents).

	Age Verification	Behavioral Detection
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	Device → vendor → platform server → centralized database.	Stays entirely on device. Nothing transmitted.
CATCHES	Minors trying to access age-restricted content.	Adults exhibiting predatory behavioral patterns.
COMPLIANCE	Triggers COPPA, GDPR, BIPA, UK AADC obligations for all users.	**No regulated data transaction occurs.**

SOURCES Gilly, T. (2026). Beyond Age Gates, Figure 1. Real Safety AI Foundation.

PART TWO

The Architecture

On-device. Deterministic. Sender-only. Zero data egress by construction.

II

Four functional components execute within the local runtime of the messaging client, architecturally constrained to run without requesting network permissions for analysis operations.

"Predatory intent is established by the offender's conduct pattern, regardless of whether the target is actually a minor."

The system analyzes only the sender's behavioral output. No data structure representing the recipient is accessed, created, or referenced by risk scoring. The constraint mirrors settled doctrine in the United States and United Kingdom, where undercover operations secure convictions on behavioral pattern alone, with adult officers posing as minors. The conduct is the offense. The target's identity is irrelevant to detecting it.

SOURCES Crown Prosecution Service, Policy Guidance on undercover stings (2019) [5]. · U.S. DOJ undercover-operation press releases [6].

Four components, all executing on-device.

01

Message Intercept

Captures sender messages from the chat pipeline. Volatile memory only; no persistent storage. No network operations.

→

02

Pattern Analysis Engine

Input normalization, behavioral-sequence FSMs, weighted rule graph, temporal pattern analyzer. Accepts sender output + local time only.

→

03

Action Engine

Visual alerts, conversation flagging, optional blocking. Optional P2P parent alert: detection category only.

↺

04

Update Mechanism

Downloads signed rule packages. Verifies signatures. Strictly unidirectional: download only. Zero telemetry.

Device boundary No data crosses it, except the optional category-only parent notification and inbound rule updates. Zero device IDs, zero telemetry.

SOURCES Gilly, T. (2026). Beyond Age Gates, Figure 2 & §4. Real Safety AI Foundation.

Deterministic by design, identical inputs yield identical outputs, every decision traceable to a rule.

Input Normalization Layer

Counters adversarial obfuscation: homoglyph mapping to canonical ASCII, Levenshtein distance against indicator dictionaries, whitespace reduction, leetspeak normalization. All deterministic.

Behavioral Sequence FSMs

Finite-state machines modeling predatory stages, initial contact, rapport, trust, isolation, boundary testing, desensitization, escalation, meeting solicitation, with deterministic transitions.

Weighted Rule Graph

A directed acyclic graph; nodes are indicators, edges are sequence relationships with fixed weights from forensic and criminological research. Composite risk by deterministic summation.

Temporal Pattern Analyzer

Tracks message frequency, time-of-day patterns, session duration, and escalation velocity with deterministic thresholds from forensic grooming-cadence literature.

SOURCES Bours & Kulsrud, Detection of Cyber Grooming in Online Conversation, IEEE WIFS 2019 [9]. · Vogt, Leser & Akbik, ACL-IJCNLP 2021 [8].

Seven conduct categories, detected from the sender alone.

Grooming Cadence
Escalating intimacy, progressive disclosure solicitation, compliment escalation, dependency building.

Isolation Tactics
Moving to private channels, discouraging disclosure to trusted adults, exclusive relational framing.

Age-Probing
Inquiries about age, grade, physical development, especially combined with other indicators.

Boundary Testing
Introduction of intimate content at incrementally increasing levels over time.

Authority & Trust
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. All categories are clinical constructs from forensic research, not keyword lists.

SOURCES Vogt, Leser & Akbik (2021) [8]. · Bours & Kulsrud, IEEE WIFS 2019 [9]. · Technology Coalition, Online Grooming research paper (2021) [11].

Rules that evolve without ever opening a way out.

Predatory tactics change, so the rule set must too, but every update channel is a potential data-egress channel, and egress is exactly what this architecture forbids. Rule definitions are stored as updateable assets, like antivirus signatures, rather than hardcoded logic. Packages are cryptographically signed; the SDK verifies signatures before ingestion.

Download only
Signed rule packages flow in. Signatures verified before ingestion.

Nothing flows back
No device identifiers, telemetry, confirmations, or user data transmitted to the update server.

The channel is strictly unidirectional. Rule evolution is unbounded; data egress remains exactly zero.

SOURCES Gilly, T. (2026). Beyond Age Gates, §4.5 (Unidirectional Definition Updates). Real Safety AI Foundation.

PART THREE

Why It Holds

Three design principles, a prior-art gap no existing system fills, and the only path that survives end-to-end encryption.

III

A comprehensive prior-art survey across patent databases, academic literature, and commercial products confirms no existing system combines all of these architectural properties.

The privacy guarantee is structural, not promised.

01

Sender-only detection

No recipient data structure is accessed, created, or referenced by risk scoring. The constraint reflects legal doctrine on conduct-based predatory intent.

02

Deterministic operation

Reproducible, auditable, resistant to adversarial examples and data poisoning. Eliminates the ethics problem of collecting exploitation material to train a model.

03

Privacy by architecture

No content transmitted, no vendor, no central database that could be subpoenaed or breached. COPPA, GDPR, and BIPA do not attach because no regulated transaction occurs.

SOURCES Gilly, T. (2026). Beyond Age Gates, §5 (Design Principles). · Chehbouni et al., arXiv:2501.12537 (2025) [10].

No identified system satisfies all eight requirements.

SYSTEM	ON-DEVICE	DETERM.	SENDER-ONLY	NO AGEVER	PLAT-AGN.	UNI-UPDATE	IN-NORM	LEGAL
Proposed	●	●	●	●	●	●	●	●
IBM 8,099,668	◐	●	○	◐	○	○	○	○
US 2021/0234823	○	○	○	○	○	○	○	○
Project Artemis	○	○	○	○	○	○	○	○
Street & Olajide	●	●	○	◐	●	○	○	○
Bark / Net Nanny	○	○	○	○	○	○	○	○

● Satisfies ◐ Partial ○ Does not satisfy

IBM comes closest, but fails on data containment, sender-only analysis, and platform generality.

SOURCES IBM U.S. Patent 8,099,668 B2 (2012) [1]. · US Patent App. 2021/0234823 A1 [2]. · Street & Olajide, ICCWS 2023 [3]. · Microsoft/Thorn Project Artemis (2020) [7].

The only detection that survives end-to-end encryption intact.

Server-side scanning is impossible under end-to-end encryption. Client-side scanning proposals create egress channels that undermine the encryption guarantee. Operating entirely on-device with no egress, this architecture is compatible with end-to-end encryption without compromising it, plausibly the only viable approach to predatory detection in encrypted messaging that does not weaken the encryption itself.

Server-side scanning	Impossible under E2EE.
Client-side scanning	Creates egress; breaks the guarantee.
On-device, no egress	Compatible. Encryption stays intact.

SOURCES EU CSA Regulation client-side-scanning proposals [11]. · Gilly, T. (2026). Beyond Age Gates, §7.3.

The same solution both sides of the safety debate could endorse.

WHO OPPOSES BIOMETRIC VERIFICATION

Civil liberties organizations, privacy

Children's advocates, biometric collection from minors

Technologists, implementation failures

Legal scholars, compliance impossibilities

WHAT THIS ARCHITECTURE REMOVES

No biometric data collected

No surveillance infrastructure created

No vendor relationships

No data transmitted off the device

Every standing objection is answered by construction, so opponents and advocates of safety mandates can land on the same side.

SOURCES Gilly, T. (2026). Beyond Age Gates, §7.2 (Political Positioning). Real Safety AI Foundation.

What this paper does not claim.

An architecture, not a product

No deployed implementation is presented. The contribution is the design and its prior-art novelty.

Precision–recall tension

Deterministic detection trades sensitivity against false positives; tuning is a genuine constraint.

Rules need expert authorship

Credible rule sets require collaboration with forensic psychologists, criminologists, and law enforcement.

Grooming patterns only

Integration needs developer cooperation, and the system does not address all online harms to minors.

SOURCES Gilly, T. (2026). Beyond Age Gates, §7.1 (Limitations). Real Safety AI Foundation.

Protecting children and protecting privacy is a false dilemma.

Age verification 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. Analyzing sender behavioral output through deterministic, on-device pattern matching identifies the conduct that constitutes the threat without any information about the recipient. No biometric data is collected. No data leaves the device. No compliance obligation is triggered. The dilemma was always an artifact of asking the wrong question from the beginning.

SOURCES Gilly, T. (2026). Beyond Age Gates. Real Safety AI Foundation.

REFERENCES

Gilly, T. (2026). *Beyond Age Gates: A Privacy-Preserving Architecture for Predatory Pattern Detection in Digital Messaging.* Real Safety AI Foundation, February 2026.

[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 Page 8 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.

REFERENCES (CONTINUED)

[11] Technology Coalition. (2021). Online grooming: Considerations for detection, response, and prevention. Research Paper. Page 9