DRAWING THE LINE

CENTER FOR ONLINE SAFETY AND LIBERTY

Drawing the Line Watchlist 2026

Platform Child Safety Governance, Due Process, and Accountability

Overview

Internet platforms have become the primary decision-makers determining what constitutes child sexual exploitation online. Through their terms of service, automated detection systems, trusted-flagger arrangements, and reporting pipelines to law enforcement, platforms now exercise powers once reserved to legislatures, courts, and police.

The Drawing the Line Watchlist 2026 will evaluate how major internet platforms govern child safety in practice—focusing not on whether child sexual abuse is taken seriously (it must be), but on how platforms draw the line between real abuse and lawful expression, and whether they do so in ways that are transparent, accountable, and consistent with due process and human rights norms.

This second edition builds on the 2025 Watchlist, which examined national legal frameworks. The 2026 edition shifts the lens to platform governance, where rules are often broader than the law, enforcement is largely automated, and errors can have devastating consequences for innocent users.

Framing the Problem

Platforms are under intense political, legal, and public pressure to detect and remove child sexual abuse material (CSAM). In response, many have adopted risk-averse, automation-heavy enforcement models that prioritise liability avoidance over accuracy, proportionality, or procedural fairness.

As a result, platforms increasingly:

- remove lawful content due to overbroad or erroneous classification under child-safety policies,
- suspend or terminate accounts without meaningful notice or explanation,
- escalate content to law enforcement without human verification, and
- offer limited or illusory appeal mechanisms.

These practices do not merely inconvenience users. They can result in false criminal suspicion, reputational harm, loss of livelihood, loss of irreplaceable personal data, re-traumatisation of survivors, and severe mental health impacts.

The central failure is not insufficient vigilance, but unaccountable vigilance. For example, platforms have:

- **Suspended or terminated accounts without meaningful human review or explanation**.
  In 2026 a man in his twenties had his Instagram account terminated after a current photo that he posted of himself wearing sports clothes was wrongly classified as a child exploitation image. Many other users have lost social media accounts and photographs in similar circumstances, with even photos of family pets triggering wrongful bans.
- **Reported innocent users to law enforcement based on automated or careless misclassification of lawful content.**
  Google reported a father to police after he used Gmail to send medical photographs of his toddler to the child's doctor. Although authorities found no wrongdoing, the user permanently lost access to his Google account and years of irreplaceable family photos and data.
- **Escalated false positives involving family or caregiving content, triggering criminal suspicion and public stigma.**
  TikTok reported an Australian grandmother to police over a non-sexual video of her massaging her infant granddaughter, sent privately to the child's mother. Media coverage falsely labelled her a child abuser before authorities confirmed no abuse had occurred.
- **Initiated law-enforcement referrals over clearly adult content, exposing users to serious legal and health consequences.**
  In ongoing U.S. litigation, a user sued Verizon and its CSAM-scanning service provider after being reported to authorities over images that were plainly adult and bore adult-site watermarks. A court has already allowed key claims to proceed, underscoring the risks of automated or negligent escalation.

These examples illustrate systemic failures of notice, explanation, human review, and accountability—precisely the governance gaps the Watchlist 2026 is designed to assess.

Project Objectives

The Drawing the Line Watchlist 2026 will:

1. **Map platform child-safety governance frameworks**
   Examine how platforms define, detect, escalate, and sanction content under child-safety rules, including the role of automation and third-party services.
2. **Evaluate due process protections for users**
   Assess whether platforms provide:
   - clear notice of alleged violations,
   - intelligible explanations of decisions,

    - meaningful human review,
    - timely and independent appeals, and
    - proportional remedies and restoration where errors occur.
3. **Assess transparency and accountability mechanisms**
   Determine what platforms disclose about:
    - detection technologies and vendors,
    - use of hash lists and classifiers,
    - relationships with reporting hotlines and law enforcement,
    - volume and error rates of child-safety enforcement actions.
4. **Identify structural drivers of over-enforcement**
   Analyse how legal pressure, regulatory threats, and reputational risk shape platform behaviour in ways that systematically increase false positives.
5. **Develop practical, rights-respecting benchmarks**
   Produce clear, actionable standards for platform governance that improve child safety outcomes without sacrificing accuracy, fairness, or trust.

Scope and Methodology

The Drawing the Line Watchlist 2026 will assess the child-safety governance practices of a representative group of major internet platforms whose services involve user-generated content, private communications, or monetisation of expressive material. Platforms will be selected to reflect diversity across:

- social networking and media-sharing services;
- messaging and communications platforms;
- hosting, publishing, and distribution services; and
- payment processors and ancillary services whose policies function as gatekeepers for online expression.

Selection will prioritise platforms whose policies or enforcement practices have demonstrable downstream effects on users' speech, livelihoods, privacy, or exposure to law-enforcement scrutiny.

Each platform will be evaluated using a structured framework built around three core dimensions:

1. **Due Process**
   Whether platforms provide users with:
    - timely and intelligible notice of enforcement actions;
    - clear explanations of the rules applied and evidence relied upon;
    - meaningful access to human review prior to severe sanctions;
    - independent and effective appeal mechanisms; and
    - proportionate remedies, including reinstatement and data restoration where errors occur.
2. **Transparency**

Whether platforms disclose:
    - child-safety rules and enforcement criteria in accessible terms;
    - the role of automated tools, classifiers, and third-party vendors;
    - relationships with reporting hotlines, trusted flaggers, and law-enforcement agencies;
    - the circumstances under which content may be escalated beyond the platform; and
    - aggregate data on enforcement volumes, error rates, and reversals.
3. **Accountability**
   Whether platforms:
    - measure and publicly report false positives and correction rates;
    - conduct internal audits or impact assessments of child-safety systems;
    - provide mechanisms for external scrutiny or expert review; and
    - demonstrate institutional learning from documented failures.

The Watchlist will draw on multiple sources of evidence, including:

- **Policy and terms analysis**, examining published rules, enforcement guidelines, and transparency reports;
- **Documented case analysis**, drawing on litigation, media reporting, regulatory filings, and first-hand accounts where corroborated;
- **Comparative assessment**, identifying common patterns and divergences across platforms;
- **Expert consultation**, engaging legal scholars, survivor advocates, technologists, and trust-and-safety practitioners to test assumptions and refine benchmarks; and
- **Human rights benchmarking**, referencing international standards on freedom of expression, privacy, procedural fairness, and child protection.

The Watchlist will not assess the moral or aesthetic value of specific content. Its focus is on process, safeguards, and governance design, rather than on individual moderation decisions in isolation.

Significance and Impact

The Watchlist 2026 addresses a critical accountability gap. While governments debate legislation, platforms are already enforcing global norms in practice, often without scrutiny or constraint.

By making platform child-safety governance legible and comparable, this project will:

- empower affected users and advocates,
- inform regulators and policymakers,
- provide constructive guidance to platforms seeking to improve, and
- strengthen public confidence that child protection systems are both effective and just.

Crucially, it reframes child safety and civil liberties as mutually reinforcing, not competing, values.

Outputs

- **Drawing the Line Watchlist 2026 report** (openly licensed)
- Platform scorecards and comparative summaries
- Briefing notes for funders, policymakers, and media
- Public presentations and targeted outreach to trust & safety communities

Funding Need

Support is sought to enable the full delivery and impact of the Drawing the Line Watchlist 2026. Funding will be used to resource the project across five interdependent areas:

1. Research and Analysis

Dedicated research and drafting capacity is required to analyse platform policies, enforcement practices, transparency disclosures, and documented case studies across multiple jurisdictions and services. This includes the systematic application of the Watchlist's due process, transparency, and accountability framework to each platform assessed.

2. Documentation of Platform Practices

Resources are needed to identify, verify, and document concrete examples of platform child-safety enforcement, including false positives, law-enforcement escalations, and appeal failures. This work is essential to ensure that the Watchlist is grounded in evidence rather than abstraction, while protecting the privacy and safety of affected individuals.

3. Expert Consultation and Peer Review

The Watchlist will draw on structured input from legal scholars, survivor advocates, technologists, and trust-and-safety practitioners. Funding will support consultation, review, and refinement of the analytical framework to ensure accuracy, balance, and credibility across disciplines.

4. Editing, Design, and Publication

Professional editing and design support are required to ensure that the Watchlist is accessible to policymakers, platform decision-makers, journalists, and civil society audiences. This includes the production of executive summaries, briefing materials, and digital publication formats.

5. Strategic Dissemination and Engagement

To maximise impact, funding will support targeted dissemination, including briefings,

presentations, and outreach to platform governance teams, regulators, and digital rights forums. Without dedicated dissemination support, there is a significant risk that the Watchlist's findings will remain confined to specialist audiences rather than informing real-world policy and practice.

About COSL

The **Center for Online Safety and Liberty (COSL)** is a nonprofit dedicated to empowering individuals and communities to thrive online by building safer digital spaces, fostering creativity, combating harm, and championing digital rights. COSL serves as an incubator for independent projects that tackle pressing issues such as age verification mandates, Section 230 rollbacks, encryption battles, and content-scanning overreach, while also developing open source trust-and-safety tools and nurturing inclusive online communities.

Contact

Center for Online Safety and Liberty
18 Bartol Street #995
San Francisco CA 94133

+1 415 650 2557
Email: info@c4osl.org
Website: https://c4osl.org

Chair: Jeremy Malcolm
jeremy@c4osl.org

Development Consultant: Sofia Bonilla
sofia@c4osl.org