

# Pedalgo Methodology and Confidence Evaluation

Risk scoring for age-inappropriate adult-to-minor contact on social platforms

A compliance-grade technical report on the scoring methodology and on the confidence that can be placed in the system's highest-scoring outputs.

SENSITIVE — TRUST & SAFETY · NOT FOR PUBLIC RELEASE

Prepared for	Regulatory and law-enforcement review	Document ID	PEDALGO-CONF-2026-06
Document owner	Pedalgo project	Version	v1.0 (draft)
Audience tier	Regulators · investigators · oversight	Date	5 June 2026

## Scope and status of this document — read first

This report specifies a **reference methodology** and a **confidence-evaluation framework** for the Pedalgo risk-scoring system. At the time of writing the production system's source code and evaluation data were **not available** to the authors for direct inspection. Consequently **every system-specific number, table and chart in this document is ILLUSTRATIVE** — synthetic values included solely to demonstrate the reporting format. They are not measurements of any deployed system and must be replaced with figures computed on real, held-out evaluation data before any operational or regulatory reliance.

Throughout, a Pedalgo score is a **risk signal that prioritises human review**. It is never, in itself, a determination that any person has committed an offence, and it must not be used for public identification, naming, or any automated adverse action.

*This document is a technical and governance specification. It is not legal advice; jurisdiction-specific obligations should be confirmed with qualified counsel.*

# 1. Executive summary

## 1.1 Purpose

Pedalgo is a triage and prioritisation system. It ranks actor accounts on a social platform by an estimated risk that the account is engaged in age-inappropriate sexualised or grooming-pattern contact with one or more likely-minor accounts. The estimate is built from two complementary sub-scores — a Deviant-Age-Contact (DAC) sub-score derived from the content and dynamics of an interaction, and a Social-Contact sub-score derived from the actor's contact pattern across the platform graph — fused into a single calibrated composite score on a 0-100 scale. The output is an ordered review queue: the highest-scoring accounts are surfaced first to trained human reviewers.

This report does two things. First, it documents the scoring methodology in full. Second — and this is the question the commissioning brief asks us to answer — it sets out how much confidence can responsibly be placed in the **top-scoring** outputs, and the evidence a reviewer, auditor, or regulator should demand before relying on them.

### What 'confidence in the top-scoring outputs' actually means

It is not a single accuracy figure. A defensible confidence statement has four parts, reported together: **(1) calibration** — does a score of 70 really correspond to a 70% posterior risk; **(2) precision@k with an uncertainty interval** — of the top k accounts the team actually reviews, what fraction are true risks, and how wide is the confidence band; **(3) subgroup stability** — does that precision hold across languages, surfaces and cohorts; and **(4) the governance firewall** — the explicit fact that the score prioritises human review and nothing adverse follows from the score alone.

## 1.2 The central question and how we answer it

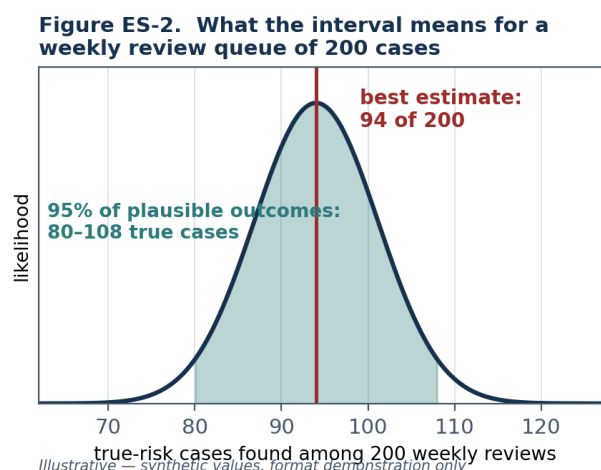
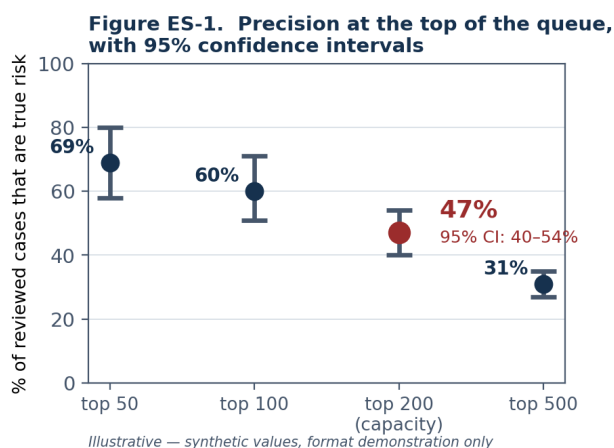
Under the conditions Pedalgo operates in, true-risk accounts are extremely rare relative to total accounts. In this severe class-imbalance regime, ordinary accuracy and even the area under the ROC curve are misleading: a model can look near-perfect on those metrics while still flooding a review queue with false positives. We therefore evaluate the system the way it is actually used — as a ranked queue under a fixed human-review capacity — using precision@k, lift, the area under the precision-recall curve, probability calibration, and explicit uncertainty intervals. Section 4 develops each of these and Section 5 presents an illustrative results set in the exact format the production system should report.

## 1.3 Headline confidence picture (illustrative)

The figures below are **illustrative** placeholders showing the format and the order of magnitude a healthy system might report; they are not measurements of Pedalgo. Their purpose is to make the confidence statement concrete.

Confidence component	Illustrative value	How to read it
Precision@k at review capacity (k = 200/wk)	0.47 (95% CI 0.40–0.54)	Of the 200 highest-scored accounts reviewed each week, roughly 94 are expected to be genuine risk cases.
Lift@200 over base rate	≈ 940×	The top-200 queue is ~940 times richer in true risk than random sampling.
Recall@200	≈ 0.47	The weekly queue captures about 47% of the true-risk accounts present that week.
Calibration error (top decile)	ECE = 0.04	High scores are slightly over-confident; corrected by temperature scaling and monitored.
Discrimination (PR vs ROC)	AUPRC ≈ 0.20 ; AUROC ≈ 0.96	ROC looks excellent; the precision-recall view is the honest one under imbalance.
Worst-cohort precision@200	0.33 (below the 0.40 floor)	At least one language cohort falls below the acceptable floor and requires a re-fit before reliance.

Figures ES-1 and ES-2 visualise the uncertainty those numbers carry. ES-1 shows the precision of the queue head with its 95% confidence interval at several review depths; ES-2 translates the operating-point interval into the range of outcomes a weekly queue of 200 reviews should expect.



Figures ES-1 (left) and ES-2 (right). The 95% confidence interval around top-of-queue precision, and what it implies for a weekly review queue of 200 cases. Illustrative — synthetic values for format demonstration.

## 1.4 The non-negotiable safeguards

- **Human-in-the-loop, always.** No suspension, report, disclosure, or other adverse action is triggered by a score. A trained reviewer is the decision-maker, and that review must be substantive, not a rubber stamp.
- **Signals, not verdicts.** A high score establishes priority for review, never guilt. False positives in this domain are gravely harmful, so the score must never leave the review team or be attached to a person's identity externally.
- **Lawful escalation only.** Where human review finds apparent exploitation, escalation follows defined legal channels — in the United States the NCMEC CyberTipline under 18 U.S.C. 2258A, internationally via ICMEC, IWF and the EU Digital Services Act trusted-flagger route — and never vigilante exposure.
- **Data protection by design.** Minors' data receives heightened protection; processing is minimised, retained only as long as justified, and fully logged for audit.
- **Auditable and contestable.** Every flag, reviewer disposition and model version is logged; affected users have a route to redress where law permits.

**Firewall principle (governs the entire system)**

Automated scoring and human judgement are separated by a hard boundary. Everything to the left of the boundary — signals, sub-scores, the composite score, the ranked queue — only **prioritises** human attention. Everything with real-world consequence happens to the right of the boundary, performed by people and lawful institutions. The score is an input to a human process, not a substitute for one.

**1.5 Bottom line for the reviewer**

Properly calibrated, capacity-matched, subgroup-audited and wrapped in the governance firewall above, a system of this design can be a high-value triage layer that materially increases the rate at which genuine risk reaches human reviewers. Equally, the same system is unsafe if its scores are read as probabilities without calibration, if precision is quoted without an uncertainty interval, if a single aggregate hides a failing cohort, or if any consequence is allowed to flow from a score without human adjudication. The remainder of this report gives a regulator the methodology and the evidentiary tests needed to tell those two situations apart.

# Contents

<b>1. Executive summary</b>	<b>2</b>
1.1 Purpose . . . . .	2
1.2 The central question and how we answer it . . . . .	2
1.3 Headline confidence picture (illustrative) . . . . .	2
1.4 The non-negotiable safeguards . . . . .	3
1.5 Bottom line for the reviewer . . . . .	4
<b>Contents</b>	<b>5</b>
<b>2. System overview and intended use</b>	<b>7</b>
2.1 Purpose and scope . . . . .	7
2.2 In-scope and out-of-scope (prohibited) uses . . . . .	7
2.3 Position in the trust-and-safety pipeline . . . . .	7
2.4 The governance firewall, stated formally . . . . .	8
<b>3. Scoring methodology</b>	<b>9</b>
3.1 Unit of analysis and signal taxonomy . . . . .	9
3.2 The Deviant-Age-Contact (DAC) sub-score . . . . .	9
3.3 The Social-Contact (network) sub-score . . . . .	9
3.4 Composite Pedalgo score and output object . . . . .	10
3.5 Handling age inference as a probabilistic signal . . . . .	10
3.6 Worked illustrative example (no sensitive content) . . . . .	10
3.7 What the methodology deliberately does not do . . . . .	10
<b>4. Confidence and calibration methodology</b>	<b>11</b>
4.1 Why default metrics mislead under rare positives . . . . .	11
4.2 Ranking metrics that matter . . . . .	11
4.3 Probability calibration . . . . .	11
4.4 Calibration diagnostics and their pitfalls . . . . .	12
4.5 Uncertainty quantification . . . . .	12
4.6 The operating point is set by human-review capacity . . . . .	13
4.7 Stating confidence in the top N rigorously . . . . .	13
<b>5. Illustrative evaluation results</b>	<b>15</b>
5.1 Precision@k with uncertainty . . . . .	15
5.2 Calibration, discrimination and the operating point . . . . .	15
5.3 Subgroup stability and fairness . . . . .	15
<b>6. Known failure modes and limitations</b>	<b>17</b>
<b>7. Fairness, bias, and subgroup audit</b>	<b>18</b>
<b>8. Governance, legal basis, and escalation</b>	<b>19</b>

8.1 Scores are signals, not determinations of guilt . . . . . 19

8.2 False-positive harms and anti-vigilantism . . . . . 19

8.3 Lawful escalation pathways . . . . . 19

8.4 The European detection-law context . . . . . 19

8.5 Data protection . . . . . 19

8.6 EU AI Act classification . . . . . 20

8.7 Documentation standards . . . . . 20

**9. Deployment-readiness recommendations 21**

**10. Limitations of this report 22**

**References 23**

**Appendix A — Glossary of metrics 25**

Appendix B — Illustrative scoring specification (pseudocode) . . . . . 25

Appendix C — Data provenance and benchmark caveats . . . . . 25

## 2. System overview and intended use

### 2.1 Purpose and scope

Pedalgo addresses one narrowly-defined trust-and-safety task: surfacing, for human review, accounts whose behaviour is consistent with an adult initiating or escalating age-inappropriate sexualised contact with a minor. It is a prioritisation instrument operating on a platform's own data, intended to sit upstream of a human moderation and child-safety workflow. It is not an adjudication system, not an identity system, and not a substitute for law-enforcement process.

### 2.2 In-scope and out-of-scope (prohibited) uses

In scope	Out of scope / prohibited
Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': 'Ranking actor accounts by estimated contact-risk to route human review capacity to the highest-risk cases first.' 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text='Ranking actor accounts by estimated contact-risk to route human review capacity to the highest-risk cases first.', textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph	Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': 'Any automated enforcement (suspension, ban, content removal) triggered by a score without human review.' 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text='Any automated enforcement (suspension, ban, content removal) triggered by a score without human review.', textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph
Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': 'Providing reviewers with the contributing signals behind a case, as context for their judgement.' 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text='Providing reviewers with the contributing signals behind a case, as context for their judgement.', textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph	Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': "Public identification, naming, 'watch-listing', or sharing of scored individuals outside the review team." 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text="Public identification, naming, 'watch-listing', or sharing of scored individuals outside the review team.", textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph
Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': 'Producing calibrated, uncertainty-bounded evidence of system reliability for oversight and audit.' 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text='Producing calibrated, uncertainty-bounded evidence of system reliability for oversight and audit.', textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph	Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': 'Treating a score as proof of intent or guilt, or presenting a score as a forensic conclusion.' 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text='Treating a score as proof of intent or guilt, or presenting a score as a forensic conclusion.', textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph
Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': "Feeding lawful escalation that is initiated by a human reviewer's finding." 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text="Feeding lawful escalation that is initiated by a human reviewer's finding.", textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph	Paragraph( 'caseSensitive': 1 'encoding': 'utf8' 'text': "Profiling minors, or retaining minors' content beyond what is necessary and lawful." 'frags': [ParaFrag(__tag__='para', bold=0, fontName='DJS', fontSize=8.2, greek=0, italic=0, link=[], rise=0, text="Profiling minors, or retaining minors' content beyond what is necessary and lawful.", textColor=Color(.101961,.121569,.168627,1), us_lines=[])] 'style': 'bulletText': None 'debug': 0 ) #Paragraph

### 2.3 Position in the trust-and-safety pipeline

Figure 1 shows where Pedalgo sits. Signals are extracted from interactions and from the contact graph; they feed two calibrated sub-scores; the sub-scores are fused into a calibrated composite with an attached uncertainty band; the composite orders a review queue. A hard governance

boundary separates this automated prioritisation from the human review and lawful escalation that follow.

**Figure 1. Pedalgo reference pipeline and the human-review firewall**

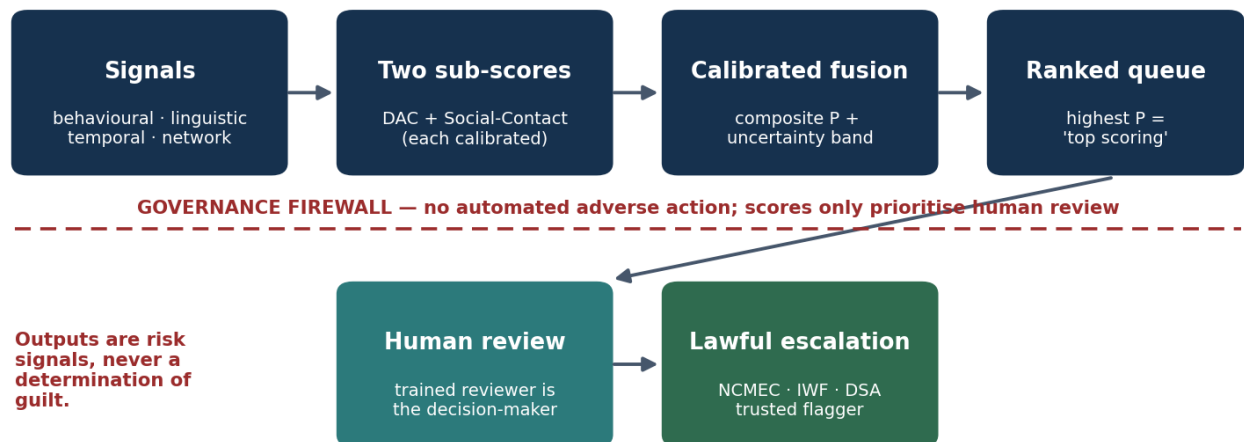


Figure 1. The Pedalgo reference pipeline. Automated components (navy) only prioritise; consequential action (teal/green) is performed by trained people and lawful institutions, on the far side of the governance firewall (dashed red).

## 2.4 The governance firewall, stated formally

- **Boundary.** A score, sub-score or ranking never causes an external effect. The only effect of a score is to change the order in which human reviewers see cases.
- **Decision authority.** The trained reviewer holds decision authority. The reviewer may act, escalate, or dismiss; the model has no authority to do any of these.
- **Escalation authority.** Escalation to NCMEC, IWF, a DSA trusted-flagger channel, or law enforcement is the output of a human finding, recorded as such, and never attributed to the model.
- **Containment.** The scored list and its contents are confidential to the review function and are not exported, published, or joined to external identity.



## 3. Scoring methodology

### How to read this section

The structure below (signal families, two sub-scores, calibrated fusion) is the reference design this report evaluates. The specific weights, thresholds and functional forms shown are **illustrative** and exist to make the method concrete and testable; the production values should be read from the system's own configuration and recorded in its model card (Section 8.7).

### 3.1 Unit of analysis and signal taxonomy

The base unit is a tuple of (actor account, target account, time window); these are aggregated to an actor-level score for queueing. Four families of signal feed the model. Crucially, the content-derived signals are model-derived probabilities over behavioural patterns, not published keyword or block lists: a literal word list would be both trivial to evade and irresponsible to circulate.

Signal family	What it captures	Feeds
Linguistic / content	Model-estimated probability of staged grooming patterns: trust-building, boundary-testing, secrecy or isolation solicitation, requests to move off-platform. Represented as learned scores, never as a keyword list.	DAC
Temporal / escalation	Trajectory and speed of topic escalation across a conversation; persistence after non-response; cross-session return. Modelled at the turn level for early detection.	DAC
Behavioural / dyadic	Initiation and reciprocity asymmetry within a dyad; who drives the conversation; response-latency patterns.	DAC + Social
Network / graph	Fan-out to many distinct likely-minor accounts (star topology); new-account burst with high minor-directed volume; off-platform migration across contacts; proximity to previously-actioned actors (advisory only).	Social

### 3.2 The Deviant-Age-Contact (DAC) sub-score

The DAC sub-score estimates, for an actor in a window, the risk that the actor is making age-inappropriate sexualised or grooming-pattern contact with a likely-minor target. Its inputs are: an estimated age gap derived from age inference on both parties (carried with its uncertainty — see 3.5); the learned grooming-pattern probability; the sexual-escalation trajectory from a turn-level early-detection model; and the dyadic initiation/reciprocity asymmetry. These are combined by a calibrated logistic model and expressed on a 0–100 scale:

$S_{DAC} = 100 \cdot \sigma( w_1 \cdot \text{age\_gap\_z} + w_2 \cdot \text{grooming\_p} + w_3 \cdot \text{escalation} + w_4 \cdot \text{asymmetry} + b )$ , with the probability passed through temperature scaling so that  $S_{DAC}/100$  approximates a true posterior. *Illustrative weights:*  $w = (0.9, 1.6, 1.2, 0.7)$ ,  $b = -3.1$ . The age-gap term enters as a standardised, uncertainty-weighted value, not as a hard rule.

### 3.3 The Social-Contact (network) sub-score

The Social-Contact sub-score estimates risk from the actor's contact pattern rather than the content of any single conversation. The dominant features are minor-fanout (the count or rate of distinct likely-minor accounts the actor initiates contact with in a window), cross-contact initiation asymmetry, new-account burst combined with high minor-directed volume, and the rate at which the actor attempts to move contacts off platform. Proximity to previously-actioned actors is included only as a low-weight, advisory feature and is explicitly fenced to avoid

guilt-by-association (3.7).

$S_{\text{Social}} = 100 \cdot \sigma( v_1 \cdot \text{minor\_fanout} + v_2 \cdot \text{init\_asymmetry} + v_3 \cdot \text{burst\_volume} + v_4 \cdot \text{offplatform\_rate} + v_5 \cdot \text{proximity\_adv} + c )$ . Illustrative:  $v = (1.4, 0.8, 0.6, 0.9, 0.2)$ ,  $c = -2.7$ . The advisory proximity weight (0.2) is deliberately the smallest and is dropped entirely in the 'conservative' operating profile.

### 3.4 Composite Pedalgo score and output object

The two sub-scores are fused in logit space, with an interaction term, and the result is re-calibrated on a held-out set (temperature or isotonic) so that the composite P approximates the actor-level posterior of true risk:

$P = 100 \cdot \sigma( \alpha \cdot \text{logit}(S_{\text{DAC}}/100) + \beta \cdot \text{logit}(S_{\text{Social}}/100) + \gamma \cdot (\text{interaction}) + b_0 )$ , recalibrated post-fusion. Illustrative:  $\alpha = 1.2$ ,  $\beta = 0.8$ ,  $\gamma = 0.3$ . A case carries the value P (0–100), the two sub-scores, the top contributing signals for reviewer context, a calibrated posterior, and an uncertainty band (a conformal set or a credible interval; see 4.5).

The ranked queue is simply the actors ordered by P. 'Top-scoring' refers to the head of this queue — the cases that will be reviewed first and on which this report's confidence question centres.

### 3.5 Handling age inference as a probabilistic signal

Age inference is a necessary but error-prone input, and Pedalgo treats it accordingly. Text-based age estimation typically carries a mean absolute error of roughly four to seven years [37], and any system that accepts a stated age, or that can be talked out of a classification by a change of writing style, is trivially gameable [36]. Pedalgo therefore (a) never uses an age estimate as a hard gate; (b) propagates the age-gap posterior, so that a case resting on an uncertain age estimate receives a correspondingly wider confidence band; and (c) prefers detection of the adult actor's behaviour, which is more robust than inferring the target's exact age. The implication for confidence is direct: cases whose rank depends heavily on an uncertain age estimate should be flagged to reviewers as lower-certainty even at the same score.

### 3.6 Worked illustrative example (no sensitive content)

Consider an abstract actor with standardised feature values  $\text{age\_gap\_z} = 1.8$ ,  $\text{grooming\_p} = 0.62$ ,  $\text{escalation} = 0.40$ ,  $\text{asymmetry} = 0.7$ . The illustrative DAC model gives a logit of  $0.9(1.8) + 1.6(0.62) + 1.2(0.40) + 0.7(0.7) - 3.1 = 0.58$ , hence  $S(\text{DAC}) \approx 100 \cdot \sigma(0.58) \approx 64$ . Suppose the network side yields  $\text{minor\_fanout} = 2.1$ ,  $\text{init\_asymmetry} = 1.0$ ,  $\text{burst} = 0.3$ ,  $\text{offplatform} = 0.8$ ,  $\text{proximity} = 0$ , giving a logit of  $1.4(2.1) + 0.8(1.0) + 0.6(0.3) + 0.9(0.8) - 2.7 = 1.94$ , so  $S(\text{Social}) \approx 87$ . Fusing:  $1.2 \cdot \text{logit}(0.64) + 0.8 \cdot \text{logit}(0.87) + 0.3 \cdot (\text{small}) + b_0$  lands the composite near  $P \approx 80$  with, say, a credible interval of  $[0.58, 0.90]$  on the posterior. This actor would enter the high-priority head of the queue, with its contributing signals shown to the reviewer — who, not the model, decides what happens next. All values here are invented for exposition.

### 3.7 What the methodology deliberately does not do

- It does not publish or rely on circulated keyword/block lists, which are evasion-prone and unsafe to distribute.
- It does not take autonomous adverse action of any kind.
- It does not treat network proximity to prior offenders as more than a small advisory nudge, to avoid guilt-by-association.
- It does not store or expose raw minor content beyond what is necessary, minimised and lawful.
- It does not output a 'verdict' field — only a priority score, sub-scores, contributing signals, and an uncertainty band.

## 4. Confidence and calibration methodology

### 4.1 Why default metrics mislead under rare positives

Pedalgo operates under severe class imbalance: genuine risk cases are a tiny fraction of all accounts. In this regime accuracy is meaningless (a model that flags nothing scores well), and even AUROC is deceptively optimistic, because the abundant easy negatives make most randomly-drawn pairs trivial to order [19][20]. Figure 4 shows the same illustrative model under both lenses: an excellent-looking ROC curve and a sobering precision-recall curve. The precision-recall view is the honest one, and it is the one a regulator should ask to see.

**Figure 4. Why AUPRC, not AUROC, governs confidence under severe class imbalance**

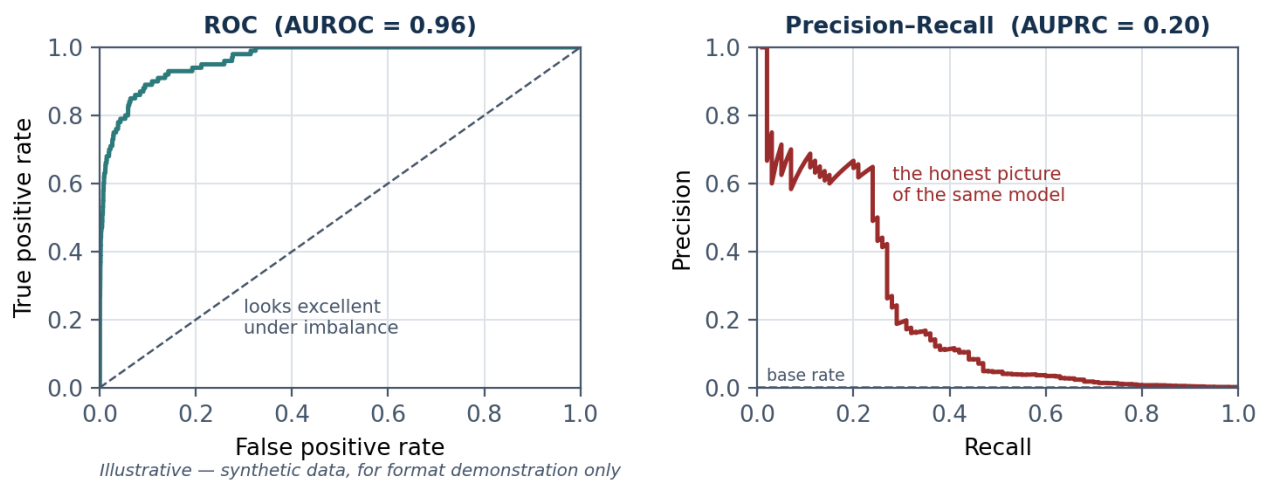


Figure 4. The same illustrative classifier scored two ways. AUROC flatters the model under imbalance; AUPRC reflects the precision a review queue would actually experience. After Davis and Goadrich [19] and Saito and Rehmsmeier [20].

### 4.2 Ranking metrics that matter

- **Precision@k** — of the top k scored accounts (k = the team's review capacity), the fraction that are true risks. This is the operational precision reviewers will live with.
- **Recall@k** — the fraction of all true-risk accounts in the period that fall in the top k; it measures coverage at the chosen capacity.
- **Lift@k** — precision@k divided by the base rate; the most intuitive way to convey to a non-technical reader how much better than random the queue is.
- **AUPRC** — area under the precision-recall curve; the primary single-number summary for model selection under imbalance [19][20]. AUROC is reported only as a supplement.

### 4.3 Probability calibration

A raw model score is almost never a probability. Calibration makes the score mean what it says: among cases scored 0.7, about 70% should be true. Boosted trees and SVMs distort probabilities in a characteristic sigmoid shape, and modern neural networks are systematically over-confident [14][15]. Four standard post-hoc methods correct this: Platt scaling (logistic remap, low-variance, best for sigmoid distortion) [14]; isotonic regression (non-parametric, more flexible but over-fits when calibration positives are scarce) [14]; temperature scaling (a single parameter on the logits, the recommended first line for neural models, preserves ranking) [15]; and beta calibration (two-parameter, captures asymmetric distortion) [16]. Pedalgo should calibrate on a **temporally held-out** set that preserves the natural positive rate, and must never calibrate on up-sampled data, which inflates the resulting probabilities.

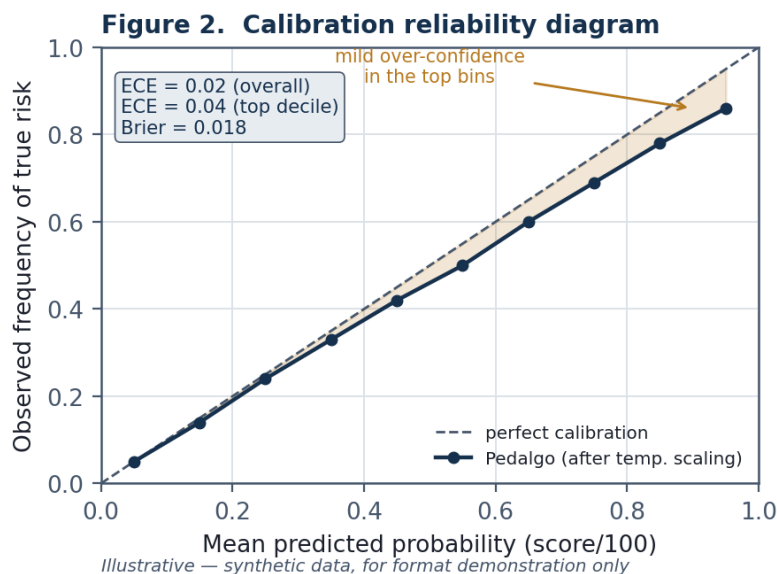


Figure 2. Illustrative reliability diagram after temperature scaling. The model is close to the diagonal but mildly over-confident in the top bins — exactly the region a top-k system relies on, so calibration is reported for the top decile separately. Method after Guo et al. [15]; diagnostics after Nixon et al. [17].

#### 4.4 Calibration diagnostics and their pitfalls

The reliability diagram (Figure 2) is the primary visual: predicted probability against observed frequency, per bin, against the diagonal. It should always be produced for the high-score tail separately, because that is where a top-k system operates and where samples are sparsest. Single-number summaries support it: Expected Calibration Error (ECE), the sample-weighted mean gap, is the most cited but is sensitive to binning and aggregates away tail problems [15][17]; Maximum Calibration Error (MCE) reports the worst bin; and the Brier score, a proper scoring rule, decomposes into uncertainty, resolution and reliability so one can see whether a good score comes from genuine discrimination or merely from the base rate [18]. Under heavy imbalance the Brier score is dominated by the negatives, so it is read alongside AUPRC, not instead of it. A crucial briefing point for executives and regulators: at a base rate of, say, 0.05%, a calibrated score of 5% is not 'low confidence' — it is a 100-fold lift over the prior and a strong, actionable signal.

#### 4.5 Uncertainty quantification

A point estimate of precision@k is not enough; the system must report how uncertain that estimate is. Four complementary techniques apply. **Bootstrap confidence intervals** resample the held-out set to put a 95% band around precision@k, and are the recommended default — with the caveat that the band is driven by the number of true positives in the test set, so the effective positive count must be reported beside it. **Conformal prediction** gives, for each case, a label set with a distribution-free coverage guarantee under exchangeability [21][22], turning a bare score into a coverage-bounded statement reviewers can act on. **Bayesian (Beta-Binomial) credible intervals** are well-suited when the test set is small, updating a prior over precision with the observed review outcomes. **Predictive entropy** (e.g. via Monte-Carlo dropout) separates cases the model finds genuinely ambiguous from cases where more data would help [23], so that intrinsically uncertain cases can be surfaced as such.

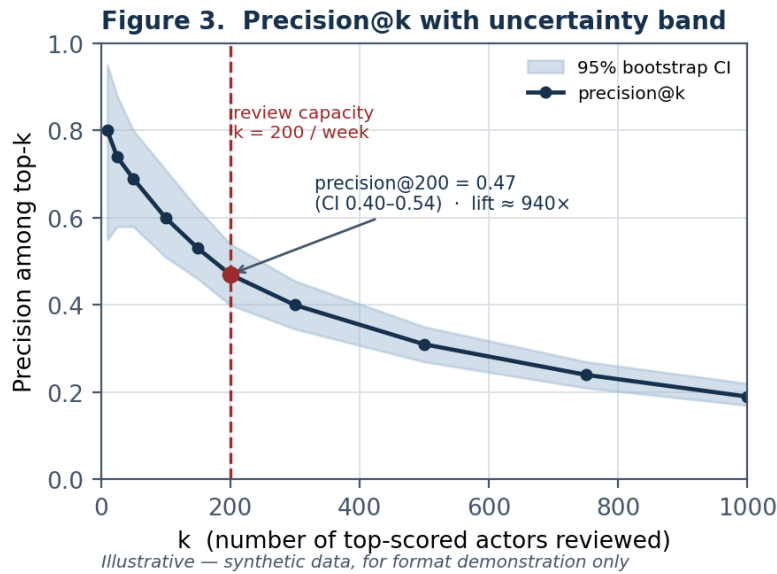


Figure 3. Illustrative precision@k with a 95% bootstrap confidence band. The band widens at small  $k$ , where few cases (and few true positives) determine the estimate. The red marker is the operating point at the team's weekly review capacity. CI methodology after Raschka [24].

#### 4.6 The operating point is set by human-review capacity

The score threshold is not a free statistical choice; it is fixed by how many cases the team can genuinely review. If capacity is  $N$  cases per week, the threshold is the score at the  $N$ th-highest rank, and every metric is then a metric at that operating point, not a global average. Raising the threshold (reviewing fewer, higher cases) raises precision but lowers recall; lowering it does the reverse, at the cost of a queue the team cannot clear. Figure 5 shows the score distribution and the capacity-set cut.

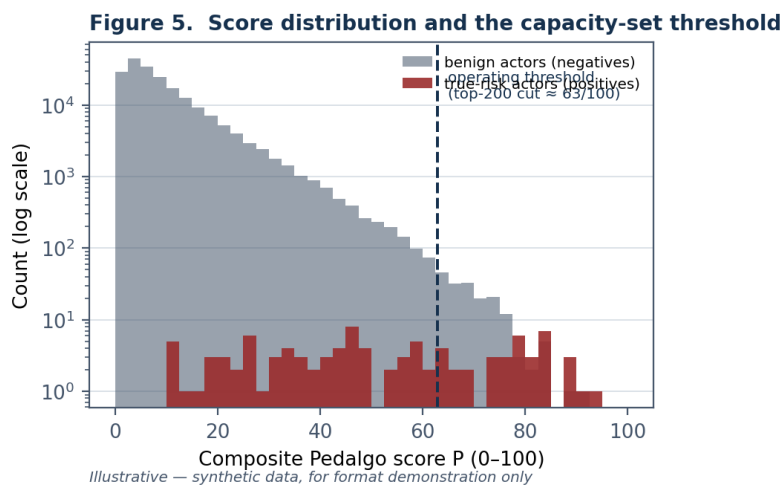


Figure 5. Illustrative score distribution (log scale). The capacity-set threshold selects the top- $k$  head of the distribution for review; the long benign tail above the threshold is the source of the false positives that human review exists to catch.

#### 4.7 Stating confidence in the top $N$ rigorously

### The four-component confidence statement (template)

"At the operating threshold corresponding to  $k = 200$  reviews per week, Pedalgo achieves  $\text{precision@200} = 0.47$  (95% bootstrap CI 0.40–0.54; 200 effective positives),  $\text{lift@200} \approx 940\times$ , and  $\text{recall@200} \approx 0.47$ . Calibration in the top decile:  $\text{ECE} = 0.04$ , with  $\sim 7\%$  over-confidence in the top bin, monitored and corrected by temperature scaling.  $\text{Precision@200}$  is stable across English and the DM surface but falls to 0.33 for one language cohort, which is below the 0.40 floor and is being re-fit. These scores prioritise human review; no adverse action follows from a score." Every number in this template is illustrative.

Metric	Definition (plain)	Why it is reported
$\text{precision@k}$	TP among the top-k scored / k	The reviewer's real hit-rate at capacity
$\text{recall@k}$	TP among top-k / all TP in period	Coverage of the true-risk universe at capacity
$\text{lift@k}$	$\text{precision@k}$ / base rate	Intuitive 'better than random' multiplier
AUPRC	area under precision-recall curve	Imbalance-honest discrimination summary
ECE / MCE	mean / max calibration gap over bins	Whether scores are trustworthy probabilities
Brier (decomp.)	mean squared error, split 3 ways	Separates discrimination from calibration
bootstrap CI	resampled interval on a metric	The uncertainty around every point estimate
conformal set	coverage-guaranteed label set	Per-case, distribution-free uncertainty

## 5. Illustrative evaluation results

### These results are illustrative

The values, tables and figures in this section are **synthetic**, generated only to demonstrate the reporting format and the relationships between metrics. They are not measurements of Pedalgo or of any deployed system. Replace every figure here with values computed on real, temporally held-out evaluation data, with the effective positive counts stated, before any reliance.

### 5.1 Precision@k with uncertainty

Table below and Figure 3 give the head of the precision@k curve with 95% bootstrap intervals. Note how the interval is wide at very small k (few cases determine it) and narrows as k grows. The operating row (k = 200) is the one quoted in the confidence statement. The illustration assumes roughly 200 true-risk actors in the scored population that week (base rate 0.0005), so cumulative true positives can never exceed 200 — an internal-consistency check any real report should also pass.

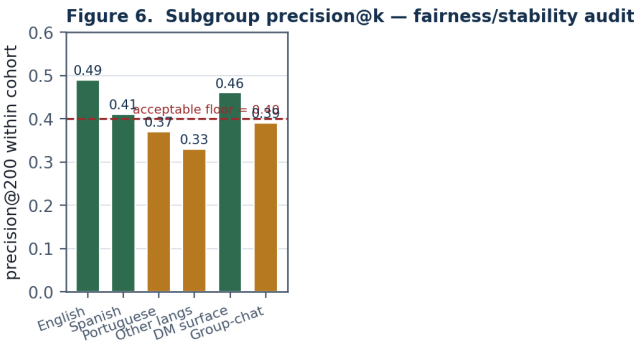
k (reviewed)	precision@k	95% CI	approx. TP in top-k	lift over base rate
10	0.80	0.55–0.95	8	1600×
50	0.69	0.58–0.80	35	1380×
100	0.60	0.51–0.71	60	1200×
200 (capacity)	0.47	0.40–0.54	94	940×
300	0.40	0.35–0.46	120	800×
500	0.31	0.27–0.35	155	620×
1000	0.19	0.17–0.22	190	380×

### 5.2 Calibration, discrimination and the operating point

Calibration (Figure 2) is close to the diagonal after temperature scaling, with mild top-bin over-confidence captured by a top-decile ECE of 0.04. Discrimination under imbalance (Figure 4) shows the characteristic gap between a flattering AUROC ( $\approx 0.96$ ) and an honest AUPRC ( $\approx 0.20$ ). The operating point (Figure 5) is set where the weekly capacity cuts the score distribution.

### 5.3 Subgroup stability and fairness

A single aggregate can hide a failing cohort. Figure 6 and the table below break precision@200 down by language and surface against an acceptable floor of 0.40. Two cohorts fall below the floor; under the governance rules of Section 7 these must be re-fit or have their thresholds adjusted before the system's output is relied upon for them.



Illustrative — synthetic data, for format demonstration only · cohorts below floor (amber) require threshold re-fit / retraining

Figure 6. Illustrative subgroup precision@200 against a 0.40 acceptable floor. Cohorts in amber are below floor and require remediation before reliance.

Cohort	precision@200	Status
English	0.49	Above floor
Spanish	0.41	Above floor (marginal)
Portuguese	0.37	Below floor — remediate
Other languages	0.33	Below floor — remediate
Direct-message surface	0.46	Above floor
Group-chat surface	0.39	Below floor — remediate



## 6. Known failure modes and limitations

The reliability of the top-scoring outputs is bounded by well-documented limitations of this whole field. A regulator should expect each to be acknowledged, quantified where possible, and mitigated.

Failure mode	Why it matters	Effect on confidence	Mitigation
Decoy-trained benchmarks	Public corpora (PJ, PAN-2012, ChatCoder2, PANC, VTPAN) derive positives from adult decoys posing as minors, not real victims [1][3][5][8].	Benchmark precision over-states real-world performance against genuine victims.	Validate on platform-internal, reviewer-confirmed cases; treat public-benchmark numbers as upper bounds.
Domain shift	Foundational data is 2012-era IRC text; modern grooming occurs on different platforms and media [9].	Top-k precision can degrade silently when deployed off-distribution.	Continuous on-distribution evaluation; monitor drift; re-fit on current platform data.
Coded / indirect language	Models miss euphemism, slang and coded speech [10].	False negatives concentrate among the most careful actors.	Red-team for coded language; ensemble with behavioural/network signals less dependent on wording.
End-to-end encryption	Message content is unavailable for server-side scanning.	Content signals vanish; reliance shifts to metadata and network signals only.	Design for a metadata-only operating profile; be explicit about its lower content sensitivity.
Age-inference error	MAE of ~4–7 years; gameable by writing style [36][37].	Cases resting on uncertain age estimates carry inflated apparent certainty.	Propagate age uncertainty into the confidence band; never gate on inferred age.
Extreme class imbalance	True positives are very rare.	Small absolute error rates still produce large false-positive volumes.	Capacity-set thresholds; report precision@k and lift, never accuracy/AUROC alone.
Single-annotator ground truth	Some benchmark labels rest on one annotator [1].	Evaluation noise, especially at the message level.	Multi-rater adjudication with inter-rater reliability reported.
Adversarial adaptation	Actors change behaviour once detection is known.	Yesterday's precision need not hold tomorrow.	Treat metrics as time-windowed; monitor for distribution change; periodic re-validation.

## 7. Fairness, bias, and subgroup audit

A model trained on historical data can learn proxies for language, dialect, region or demographic group, which in this domain creates both a fairness harm and legal exposure. The audit obligations below are part of the system, not an afterthought, and their results condition whether the output may be relied upon for a given cohort [33].

- **Subgroup precision and recall.** Compute precision@k, recall@k and AUPRC separately for each major language, surface, region and cohort (Figure 6). Define an acceptable floor; a cohort below floor is not relied upon until remediated.
- **Differential calibration.** Produce reliability diagrams per cohort: a model well-calibrated overall can be mis-calibrated for a specific community, systematically mis-ranking its cases.
- **Feature attribution.** Identify the features driving top-scored cases and flag any acting as demographic proxies (regional slang, language style) for removal or regularisation.
- **Threshold adjustment.** Where precision differs by cohort, adjust per-cohort operating thresholds rather than imposing one global cut that fails a subgroup.
- **Contestability and redress.** Where law permits, provide affected users an explanation and an appeal route; GDPR Articles 13–15 inform what must be made available [31].

### Why this is also a legal requirement

Under the GDPR, processing must be fair and transparent and a decision with significant effect may not rest solely on automated processing (Article 22) [31]; under the EU AI Act a system of this kind is likely high-risk and subject to bias-management and oversight obligations [32]. A documented subgroup audit is part of meeting both.

## 8. Governance, legal basis, and escalation

### 8.1 Scores are signals, not determinations of guilt

No score, however high, is evidence that a person has committed an offence. The score prioritises human review; the reviewer is the decision-maker; and even the reviewer's finding is an operational judgement, not a legal verdict, which belongs to the courts. Where automated processing has a significant effect on a person, the GDPR requires that the decision not be solely automated and that any human involvement be **meaningful** rather than a rubber stamp [31]. Pedalgo's human-review step is designed to meet that bar: reviewers are trained on the model's limitations and base rates and record substantive dispositions.

### 8.2 False-positive harms and anti-vigilantism

In this domain a false positive is not a minor inconvenience: wrongly associating an adult with child exploitation can destroy reputation, employment and family, and can provoke vigilante harm. The system design therefore hard-codes that the scored list never leaves the review team, is never attached to a public identity, and never justifies naming or exposure. Any public 'watch-list' use is prohibited (Section 2.2).

### 8.3 Lawful escalation pathways

When human review concludes a case shows apparent child sexual exploitation, escalation follows defined legal channels and is never improvised. The principal mechanisms:

Jurisdiction / body	Mechanism	Basis / note
United States — NCMEC	CyberTipline report of apparent CSE	Mandatory ESP reporting under 18 U.S.C. 2258A; scope expanded by the REPORT Act (2024). 20.5M reports in 2024 [25][26].
International — ICMEC	Coordination across 120+ countries	Routes and supports CyberTipline-equivalent reporting across jurisdictions [27].
United Kingdom — IWF	Hash lists and assessed-content reporting	Operates PhotoDNA-based image-hash lists; an EU DSA trusted flagger [28].
European Union — DSA	Trusted-flagger priority notice	Article 22 DSA: priority handling of notices from designated trusted flaggers [29].

#### The escalation is a human output

Each escalation is recorded as the output of a reviewer's finding, with the model version and contributing signals logged for audit — but the escalation is attributed to the human decision, never to the score.

### 8.4 The European detection-law context

The EU legal basis for voluntary detection in private communications has been in flux. The interim ePrivacy derogation that permitted voluntary CSAM detection in interpersonal communications reached its extended expiry in April 2026, and at the time of writing the proposed CSA Regulation (widely debated as 'chat control') remains unresolved, leaving a contested gap that bodies such as the IWF have described as a child-safety emergency [30]. This is a fast-moving area: any EU deployment must confirm the current legal basis with counsel before processing message content, and should be designed to fall back to a metadata-only profile where content scanning lacks a lawful basis.

### 8.5 Data protection

- **Principles (GDPR Art. 5).** Lawfulness, fairness, transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality [31].
- **Children's data (Art. 8).** Minors' data receives heightened protection — restricted access, stronger minimisation, and short, justified retention [31].
- **Automated decisions (Art. 22).** Routing to meaningful human review keeps the system out of the 'solely automated' prohibition; the human step must be substantive [31].
- **Auditability.** Every flag, reviewer disposition, escalation, and model version is logged with timestamp and reviewer ID, enabling regulatory inspection and inter-rater analysis.

## 8.6 EU AI Act classification

A child-safety risk-scoring system used to inform law-enforcement-adjacent action is likely to fall within the EU AI Act's high-risk categories, bringing obligations on risk management, data governance, transparency, human oversight, accuracy and robustness, and post-market monitoring [32]. The evaluation discipline in this report — calibration, uncertainty, subgroup audit, documentation — is largely what conformity with those obligations requires in practice.

## 8.7 Documentation standards

- **Model card** [34] — intended use (prioritisation for human review only), out-of-scope uses, training-data provenance and vintage, per-subgroup evaluation, calibration method, known limitations.
- **Datasheet for the dataset** [35] — source of positive labels, time window, and known population shifts between training and deployment.
- **System card / impact assessment** — the whole pipeline, human workflow, escalation triggers and data flows, aligned to the AI Act conformity assessment and DSA audit expectations [32].

## 9. Deployment-readiness recommendations

The following are the concrete conditions under which the top-scoring outputs can be relied upon. They are written so an auditor can check each as present or absent.

#	Recommendation	Done when
1	Calibrate scores on a temporally held-out set (temperature first line; isotonic/beta as needed).	Top-decile ECE reported and within target; reliability diagram archived.
2	Report precision@k with a 95% bootstrap CI at the true review capacity, with effective positive counts.	Confidence statement (4.7) produced each cycle.
3	Run the subgroup audit with an explicit floor; do not rely on below-floor cohorts.	Per-cohort table and per-cohort calibration published; remediation tracked.
4	Log every flag, disposition, escalation and model version.	Immutable audit log queryable by oversight.
5	Operate a documented human-review SOP with reviewer training on limitations and base rates.	SOP signed off; reviewer calibration measured (inter-rater reliability).
6	Operate a documented lawful-escalation SOP (NCMEC / IWF / DSA), human-initiated only.	Escalations attributable to human findings, with legal basis recorded.
7	Publish model card, datasheet and system card; map to AI Act / DSA duties.	Documents version-controlled alongside the model.
8	Commission independent evaluation / audit before high-stakes reliance.	External report received and actioned.
9	Red-team coded language and adversarial adaptation; schedule periodic re-validation.	Red-team findings logged; re-validation cadence defined.
10	Confirm the current legal basis for any content processing (esp. EU) with counsel.	Written legal basis on file; metadata-only fallback specified.

### One-line readiness test

The top-scoring outputs are fit to rely upon when, and only when, they are **calibrated**, quoted with an **uncertainty interval at the real review capacity, stable across audited subgroups**, and wrapped in a **human-review-and-lawful-escalation firewall**. Absent any one of these, the queue may still be useful internally, but its outputs must not drive consequence.

## 10. Limitations of this report

- **System artefacts unavailable.** The production system's source code and evaluation data were not available to the authors, so the methodology here is a reference specification and all system-specific numbers are illustrative. Supplying the methodology, configuration, or an evaluation export would let these be replaced with real values.
- **Illustrative figures.** Every metric, table and chart attributed to Pedalgo is synthetic and for format demonstration only; figures use independent synthetic datasets and are not mutually calibrated.
- **Literature currency.** The evidence base is current to mid-2026; the EU legislative position in particular is moving and must be re-checked at deployment.
- **Not legal advice.** Jurisdiction-specific obligations (mandatory reporting, lawful basis, retention) must be confirmed with qualified counsel.

### To turn this into a measured report

Provide (a) the system's methodology and configuration, and (b) an evaluation export — scored cases with reviewer-confirmed labels, ideally time-stamped and with language/surface tags. With those, the illustrative numbers in Sections 1, 4 and 5 can be replaced by real, calibrated, uncertainty-bounded results, and the subgroup audit can be run for your actual cohorts.

## References

- [1] Inches, G. and Crestani, F. (2012). Overview of the International Sexual Predator Identification Competition at PAN-2012. CEUR Workshop Proceedings, Vol. 1178. <https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-InchesEt2012.pdf>
- [2] PAN at CLEF 2012 — Sexual Predator Identification task (Webis). <https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html>
- [3] Kontostathis, A., West, W., Garron, A., Reynolds, K. and Edwards, L. (2012). Identifying Predators Using ChatCoder 2.0. Notebook for PAN at CLEF 2012, CEUR-WS. <https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-KontostathisEt2012.pdf>
- [4] McGhee, I. et al. (2011). Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce*, 15(3), 103–122.
- [5] Vogt, M., Leser, U. and Akbik, A. (2021). Early Detection of Sexual Predators in Chats. ACL-IJCNLP 2021, 4985–4999. <https://aclanthology.org/2021.acl-long.386>
- [6] Early Sexual Predator Detection datasets (PANC, VTPAN, preprocessing). GitLab. <https://gitlab.com/early-sexual-predator-detection/eSPD-datasets>
- [7] Rebedea, T. (2017). Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, 23(4), 589–616. <https://doi.org/10.1017/S1351324916000395>
- [8] Leiva-Bianchi, M. et al. (2025). Effectiveness of machine learning methods in detecting grooming: a systematic meta-analytic review. *Scientific Reports*, 15, 9008. <https://doi.org/10.1038/s41598-024-83003-4>
- [9] An, H. et al. (2025). Toward Integrated Solutions: A Systematic Interdisciplinary Review of Cybergrooming Research. arXiv:2503.05727. <https://arxiv.org/abs/2503.05727>
- [10] Ringenber, T. R. et al. (2025). A Fuzzy Evaluation of Sentence Encoders on Grooming Risk Classification. arXiv:2502.12576. <https://arxiv.org/abs/2502.12576>
- [11] Chehbouni, K. et al. (2025). Enhancing Privacy in the Early Detection of Sexual Predators Through Federated Learning and Differential Privacy. AAAI 2025 / arXiv:2501.12537. <https://arxiv.org/abs/2501.12537>
- [12] Terre des Hommes Netherlands (2013). Webcam Child Sex Tourism (the 'Sweetie' operation). <https://www.terredeshommes.nl>
- [13] van der Hof, S. et al. (eds.) (2019). Sweetie 2.0: Using Artificial Intelligence to Fight Webcam Child Sex Tourism. T.M.C. Asser Press / Springer. ISBN 978-94-6265-288-0.
- [14] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. ICML 2005, 625–632. <https://doi.org/10.1145/1102351.1102430>
- [15] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. ICML 2017. arXiv:1706.04599. <https://arxiv.org/abs/1706.04599>
- [16] Kull, M., Silva Filho, T. and Flach, P. (2017). Beta Calibration. AISTATS 2017, PMLR 54, 623–631. <https://proceedings.mlr.press/v54/kull17a.html>
- [17] Nixon, J. et al. (2019). Measuring Calibration in Deep Learning. CVPR Workshops 2019. arXiv:1904.01685. <https://arxiv.org/abs/1904.01685>
- [18] Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1–3.
- [19] Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. ICML 2006, 233–240.
- [20] Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3). <https://doi.org/10.1371/journal.pone.0118432>
- [21] Angelopoulos, A. N. and Bates, S. (2021/2023). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511. <https://arxiv.org/abs/2107.07511>
- [22] Sadinle, M., Lei, J. and Wasserman, L. (2019). Least Ambiguous Set-Valued Classifiers with Bounded Error Levels. *JASA*, 114(525), 223–234. <https://arxiv.org/abs/1609.00451>
- [23] Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ICML 2016. arXiv:1506.02142. <https://arxiv.org/abs/1506.02142>
- [24] Raschka, S. (2022). Creating Confidence Intervals for Machine Learning Classifiers; see also arXiv:2406.08099. <https://sebastianraschka.com/blog/2022/confidence-intervals-for-ml.html>
- [25] National Center for Missing and Exploited Children (2025). CyberTipline 2024 Data; REPORT Act (Pub. L. 118-79, 2024). <https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata>
- [26] Thorn (2025). What the 2024 NCMEC CyberTipline Report Says About Child Safety. <https://www.thorn.org/blog/what-the-2024-ncmec-cybertipline-report-says-about-child-safety/>

- [27] International Centre for Missing and Exploited Children. CyberTipline (NCMEC) resource. <https://www.icmec.org/>
- [28] Internet Watch Foundation. Image Hash List / PhotoDNA. <https://www.iwf.org.uk/our-technology/our-services/image-hash-list/>
- [29] Article 22 Digital Services Act: Trusted Flaggers — Internet Policy Review (2025); European Commission guidance. <https://policyreview.info/articles/analysis/article-22-digital-services-act>
- [30] ePrivacy derogation expiry and the EU CSA Regulation debate (2026): Digital Watch Observatory; Freshfields; European Parliament press room; IWF. <https://dig.watch/>
- [31] Regulation (EU) 2016/679 (GDPR), Articles 5, 8, 13–15, 22. <https://gdpr-info.eu/>
- [32] Regulation (EU) 2024/1689 (Artificial Intelligence Act), high-risk systems, Annex III. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [33] Barocas, S., Hardt, M. and Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. <https://fairmlbook.org/>
- [34] Mitchell, M. et al. (2019). Model Cards for Model Reporting. ACM FAccT 2019, 220–229. arXiv:1810.03993. <https://arxiv.org/abs/1810.03993>
- [35] Gebru, T. et al. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- [36] Age Verification Providers Association (2026). AI Needs More Than Age Inference to Protect Kids. <https://avpassociation.com/>
- [37] Klein, A. Z. et al. (2022). ReportAGE: automatically extracting the exact age of Twitter users from self-reports. PLOS ONE. <https://doi.org/10.1371/journal.pone.0262087>



## Appendix A — Glossary of metrics

Term	Plain-language definition
Base rate	The fraction of all accounts that are genuinely true-risk; here, very small.
precision@k	Among the k highest-scored accounts, the share that are true positives.
recall@k	Among all true-risk accounts in the period, the share that land in the top k.
lift@k	precision@k divided by the base rate — how many times richer the queue is than random.
AUROC	Area under the ROC curve; over-optimistic under heavy imbalance.
AUPRC	Area under the precision-recall curve; the imbalance-honest discrimination summary.
Calibration	The property that a score of p corresponds to a true probability of about p.
ECE / MCE	Expected / maximum calibration error — mean / worst gap between score and observed frequency.
Brier score	Mean squared error of probabilistic predictions; decomposes into uncertainty, resolution, reliability.
Temperature scaling	A one-parameter calibration of a model's logits; first-line fix for over-confident neural nets.
Bootstrap CI	A confidence interval obtained by resampling the evaluation set.
Conformal prediction	A method giving each case a label set with a guaranteed coverage probability.
Predictive entropy	A per-case uncertainty measure separating ambiguous cases from data-poor ones.

## Appendix B — Illustrative scoring specification (pseudocode)

```
# Illustrative only — replace weights with the system's configured values.
def dac_subscore(age_gap_z, grooming_p, escalation, asymmetry):
    z = 0.9*age_gap_z + 1.6*grooming_p + 1.2*escalation + 0.7*asymmetry - 3.1
    return 100*sigmoid(temperature_scale(z))

def social_subscore(minor_fanout, init_asym, burst, offplatform, proximity_adv):
    z = 1.4*minor_fanout + 0.8*init_asym + 0.6*burst + 0.9*offplatform + 0.2*proximity_adv - 2.7
    return 100*sigmoid(z)

def composite(s_dac, s_social):
    z = 1.2*logit(s_dac/100) + 0.8*logit(s_social/100) + 0.3*interaction(s_dac, s_social) + b0
    P = 100*sigmoid(recalibrate(z)) # temperature or isotonic on held-out set
    return P, credible_interval(z) # value + uncertainty band

# Output per case: P, s_dac, s_social, top_signals, posterior, uncertainty_band
# Effect of output: change review-queue ORDER only. No automated adverse action.
```

## Appendix C — Data provenance and benchmark caveats

Dataset / source	Positive-label provenance	Principal caveat
Perverted-Justice corpus	Adult decoys posing as minors	Not real victims; English/US; pre-2015 [1][3]
PAN-2012	PJ positives + IRC/Omegle negatives	~1% positives by design; 2012-era IRC text [1]
ChatCoder2 (CC2)	497 PJ conversations, all positive	No internal negatives; small; decoy-based [3]
PANC / VTPAN	Derived from PAN-2012 + CC2	Inherits decoy and vintage limitations [5][6]
Platform-internal cases	Reviewer-confirmed dispositions	The only realistic basis for deployment metrics; must be audited for label quality

### Final reminder

Nothing in this document is a measurement of a deployed system. It is a methodology and a confidence-reporting framework. The Pedalgo score prioritises human review; it is never a determination of guilt, and it must never be used for public identification or any automated adverse action.