Pedalgo in plain language

What an evaluation of 10,000 users would show — and how sure we'd be about the top 10

The short version

Pedalgo gives every account a risk score from 0 to 100. The score reflects two things, combined: **how** an account talks to likely minors, and **how many** likely minors it tries to reach. The highest-scoring accounts go to trained people for review. The score never punishes anyone by itself.

Imagine a platform with **10,000 users**, and suppose **5** of them are genuine risks. If reviewers check the **10 highest-scoring accounts**, a well-built system would put about **4 of the 5 real cases** in that top 10. So: about 4 of the 10 flags are real, about 6 are false alarms — and that is exactly why humans review every case before anything happens.

The five numbers that matter	Value	Meaning
Users scored	10,000	Everyone gets a score; almost everyone is never looked at.
Genuine risks hidden among them	5	Real cases are very rare — that rarity drives everything.
Cases humans review	top 10	Review effort goes where the risk is concentrated.
Real cases found in the top 10	about 4	Best estimate. 6 of 10 flags will be false alarms.
The honest range (confidence)	2 to 7	94% of the time the true number lands in this range.

The one rule to remember

A high score means **look here first** — it never means **guilty**. People decide, the computer only sorts. Confirmed cases go to the proper authorities, never to public exposure.

Illustrative numbers

Every number in this explainer is invented to show how the evaluation works. Real values must come from a real evaluation of the live system. Companion document: *Pedalgo — Methodology and Confidence Evaluation* (PEDALGO-CONF-2026-06).

The problem: 5 needles in a 10,000-user haystack

Genuine predatory accounts are very rare. That is good news for the platform — and the central difficulty for any detection system. When only 1 user in 2,000 is a real risk, even a tiny error rate, applied to 10,000 people, produces more false alarms than true finds.

Figure 1. The problem: find 5 people hidden among 10,000

Illustrative — invented numbers to show the format, not real measurements

Rarity also breaks the most familiar statistic. A system that flags **nobody** is right about 9,995 of 10,000 users — **99.95% 'accurate'** — while catching no one. So Pedalgo is never judged on accuracy. It is judged on what happens at the top of its ranking: of the cases it pushes to the front of the queue, how many are real?

> **Why this matters to you as a reader**
> Whenever someone quotes a single impressive-sounding number for a system like this, ask two questions: **'Of the cases it flags, how many are real?'** and **'How wide is the uncertainty?'** Those two questions are what the rest of this document answers.

What the evaluation shows

The evaluation scores all 10,000 users, ranks them, sends the top 10 to human reviewers, and then checks the outcome of every review. One picture and one matrix summarise everything:

Figure 2. What happens to 10,000 users

A score never punishes anyone by itself. Action only follows human review — and confirmed cases go to the proper authorities, never to public exposure.

Illustrative — invented numbers to show the format, not real measurements

The results matrix. Every user falls into exactly one of four boxes. Counts below are the expected outcome for the illustrative scenario:

	Truly risky (5 people)	Not risky (9,995 people)
Flagged top 10	**4** correct catches — real cases found (true positives)	**6** false alarms — cleared by human review, no action (false positives)
Not flagged 9,990 users	**1** missed — a real case below the top 10 (false negative)	**9,989** correctly left alone (true negatives)

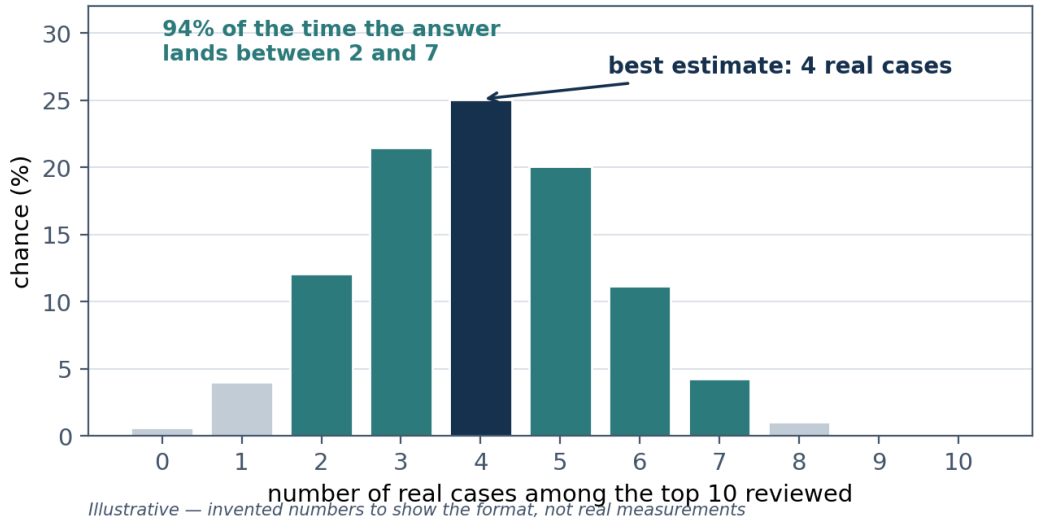
The matrix adds up: 4 + 6 = 10 flagged; 4 + 1 = all 5 real cases; 10 + 9,990 = 10,000 users. Illustrative — invented numbers to show the format.

What we measure	Result	In plain words
Hit rate of the top 10 (precision)	4 of 10 = 40%	When the system points, it is right a bit less than half the time.
Coverage of real cases (recall)	4 of 5 = 80%	Most real cases make it into the reviewed top 10.
False alarms among the flags	6 of 10	The reason no score ever acts on its own.
Missed real cases	1 of 5	Sits below the cut this week; further evidence accumulates.
'Accuracy'	99.93%	Sounds great, means nothing here — flagging no one scores 99.95%.

How confident are we about the top 10?

'About 4 of the 10 are real' is a best estimate, not a promise. With numbers this small, chance plays a big role: run the same week twice and you would not get the same result. The honest statement is a **range**, and the chart shows it.

Figure 3. How many of the top 10 will be real? The honest answer is a range



Illustrative — invented numbers to show the format, not real measurements

How to read it: if the system's true hit rate is 40%, then in **94 weeks out of 100** the top 10 will contain **between 2 and 7 real cases**. Finding only 2 in a given week does not mean the system broke, and finding 7 does not mean it suddenly improved. Single weeks prove little either way.

Why the range is so wide — because both numbers involved are tiny: 10 reviews, 5 real cases. Small samples always produce wide ranges. This is a property of arithmetic, not a flaw in the system.

- **Review more cases.** A top-50 queue gives the estimate more room to settle (at the cost of more false alarms to clear).
- **Pool more weeks.** Twenty weeks of top-10 reviews is 200 data points — the range narrows roughly with the square root of the data.
- **Check calibration.** When the system says '80', it should be right about 80% of those times. That is tested separately, and it is what makes the scores themselves trustworthy.

> **What a regulator should ask for**
> Not 'the accuracy', but: the hit rate of the reviewed queue **with its range**, measured over enough weeks, broken down by language and platform surface, plus proof that no action ever followed from a score without human review.

The rules that keep this safe

- **1. People decide.** A score changes the order of the review queue. It triggers nothing on its own.
- **2. Scores stay inside.** The ranked list never leaves the safety team and is never tied to a public identity.
- **3. No naming, ever.** Public exposure or 'watch-lists' are prohibited uses — a false accusation in this domain destroys an innocent life.
- **4. Confirmed cases go to the law.** Escalation is to lawful bodies (in the US the NCMEC CyberTipline; the IWF in the UK; EU trusted-flagger channels) — initiated by a human finding.
- **5. Fairness is checked.** Hit rates are measured per language and per community; a group with worse results gets fixed, not ignored.
- **6. Everything is logged.** Every flag, every review decision, every escalation, every model version — auditable end to end.

How to read a Pedalgo score

Score band	What it means	What happens
90–100	Highest priority	Reviewed first, same day.
70–89	Elevated	Reviewed within the week.
40–69	Noted	Queued; reviewed as capacity allows.
0–39	No signal	Nothing happens.

In every band the rule is identical: **review first, act only after human judgement, escalate only through lawful channels.**

To make these numbers real
Provide an evaluation export — scored cases with reviewer-confirmed outcomes, over as many weeks as possible, tagged by language and surface. The matrix on page 3 and the range on page 4 can then be recomputed with real data, and this explainer reissued with measured values.

Final reminder
Everything here is an illustration of how Pedalgo's evaluation works — not a measurement of a deployed system. And in deployment as on paper: a score prioritises human review; it is never a determination of guilt.