How many would Pedalgo flag in Sweden?

A worked estimate for Facebook — accounts flagged, and how many are likely real

PLANNING ESTIMATE · SENSITIVE — TRUST & SAFETY

Question	Of Facebook accounts in Sweden, how many would be flagged, and how many are real?
Method	Transparent low / base / high estimate; sourced population + illustrative detection rates
Date	5 June 2026 · PEDALGO-SE-2026-06 · v1.0

What this is — and is not

This is an order-of-magnitude **planning estimate**, not a measurement. **Population and usage figures are real and sourced** (Statistics Sweden; DataReportal/Meta). **The detection rates are illustrative** — taken from the Pedalgo reference model, because no evaluation of a live system was available. The single most uncertain number — how many genuine offenders actually exist — is shown as a wide range, and the real answer could fall outside it.

A flagged account is **a prompt for human review, never a finding of guilt**. As the numbers below show, most flagged accounts are false alarms — which is exactly why a person, not the score, decides.

Companion documents: Pedalgo — Methodology and Confidence Evaluation (PEDALGO-CONF-2026-06) and Pedalgo in plain language (PEDALGO-EXPL-2026-06).

The estimate, in one page

Sweden has about **10.6 million** people and roughly **5.8 million adult Facebook accounts**. How many Pedalgo would "flag" is not a fixed number — it is a dial the operator sets. Flag only the very top of the ranking and almost every flag is real, but you catch few cases; flag more widely and you catch more, but most flags become false alarms that human review must clear. The table shows the trade-off for the base case.

If reviewers look at…	Flagged	Likely real	False alarms	% of flags real	Share of real cases caught
the top 10 accounts	**10**	**~10**	**~0**	**95%**	**under 1%**
the top 100	**100**	**~75**	**~25**	**75%**	**~3%**
the top 1,000	**1,000**	**~340**	**~660**	**34%**	**~12%**
the top 2,900 (0.05%)	**2,900**	**~570**	**~2,330**	**20%**	**~20%**
the top 10,000	**10,000**	**~940**	**~9,060**	**9%**	**~32%**

The headline

On the base assumptions, an estimated **~2,900 genuine risk accounts** exist among Sweden's adult Facebook users (plausible range **~1,200 to ~16,000**). If reviewers work a shortlist of the **top ~2,900** highest-scoring accounts, roughly **570 turn out to be real** and about **2,330 are false alarms** — and that effort catches only about **one in five** of the real cases. Reviewing only the **top 100** is far more accurate (about 75 of 100 real) but catches far fewer.

Figure 1. From 10.6 million people to a reviewable shortlist

Only ~30,000–95,000 of Sweden's ~635,000 teenagers use Facebook at all — so contact opportunities on Facebook specifically are limited; most teen risk has moved to other apps. A flag is a prompt to review, never a verdict.

Population figures sourced (SCB, DataReportal). Detection rates illustrative.

How the estimate is built

The calculation has three honest ingredients: how many adult accounts there are (well known), how many of them are genuine risks (barely known), and how good the ranking is (assumed). We carry a low, base and high value for each uncertain input so the answer is a range, not a false point.

Input	Low	Base	High	Source / status
Sweden population	—	**10.6 M**	—	Statistics Sweden, end-2025 [1]
Adults 18+ on Facebook	**5.8 M**	**5.8 M**	**8.2 M**	DataReportal/Meta; NapoleonCat [2][3]
Teens 13–17 in Sweden	—	**~635,000**	—	SCB age structure (derived) [1]
Teens 13–17 on Facebook	**~30,000**	**~64,000**	**~95,000**	Inferred; teens skew to other apps [4][5]
Genuine risk accounts (rate)	**0.02%**	**0.05%**	**0.20%**	ILLUSTRATIVE; anchored to [8][9][10]
…which is, in people	**~1,200**	**~2,900**	**~16,000**	= rate × adult accounts
Ranking quality (AUROC)	**0.94**	**0.96**	**0.98**	ILLUSTRATIVE (reference model)

The base rate is the weak link

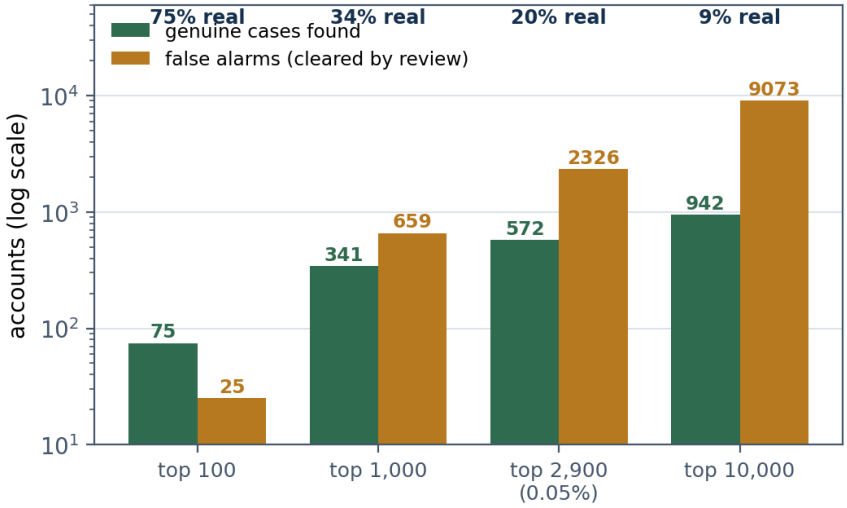
Nobody knows how many adults are actively making age-inappropriate contact on Facebook in Sweden. Our illustrative 0.05% is anchored to real signals — Meta disabled about **600,000 predatory accounts** in a single 2023 sweep (~0.018% of its users) [9]; surveys suggest ~1% of men report sexual interest in minors [10] — but only a fraction act, on this platform, in a detectable way. Treat the 'real cases' figures as order-of-magnitude only.

Everything else is arithmetic. For any shortlist size, the ranking model implies how many of the flagged accounts are genuine (precision) and what share of all real cases that captures (recall). The detection model, thresholds and confidence method are documented in the companion methodology report.

How many flagged — and how many real

Figure 2 makes the trade-off visible. Green bars are genuine cases found; amber bars are false alarms that human review clears. Note the log scale: at the top 100, three in four flags are real; by the top 10,000, that has fallen below one in ten while the false-alarm pile has grown into the thousands.

Figure 2. The dial: flag fewer for higher accuracy, or more for wider coverage



Illustrative — base case (≈2,900 genuine cases assumed to exist). Population figures sourced (SCB, DataReportal). Detection rates illustrative.

There is no single "right" setting. A small, high-accuracy shortlist respects reviewers' time and minimises the number of innocent people examined, but leaves most real cases for later. A larger shortlist finds more, at the cost of a heavy false-alarm load — every one of which is a real person who must be cleared, fairly and quietly, by a human. The right dial is a policy choice about review capacity and the acceptable cost of false alarms, made in the open and logged.

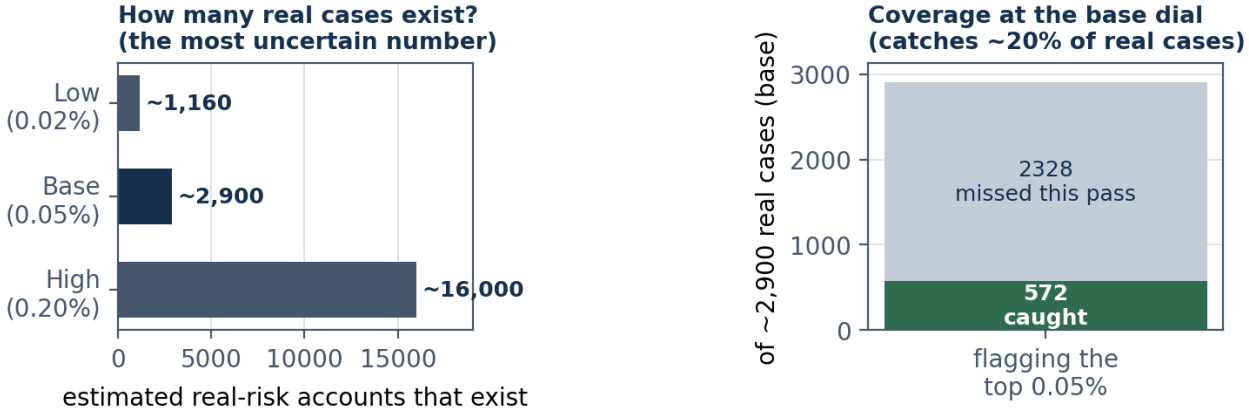
Why the percentages look low — and why that is honest

Genuine cases are rare (well under one in a thousand accounts). When you search for something that rare, even an excellent ranking returns many false alarms. That is arithmetic, not failure — and it is the whole reason a score may never act on its own.

How many real cases exist, and how many we would miss

"Flagged" and "real" are different questions from "how many exist". The number that actually exist depends only on the base rate, and that is the most uncertain input of all. Figure 3 (left) shows the spread: from roughly 1,200 in the low case to about 16,000 in the high case, with a base of about 2,900. Figure 3 (right) shows that even at the base operating point, a single pass over the top 0.05% catches only about a fifth of them — the rest sit below the cut until further evidence accumulates.

Figure 3. Real cases that exist, and how many one pass catches



Illustrative + sourced base rates. Range reflects deep uncertainty in prevalence.

Two reasons the true total is hard to pin down

First, prevalence estimates for offending span an order of magnitude. Second, Facebook is **not** where most Swedish teenagers are — only tens of thousands of 13–17s use it, versus hundreds of thousands on TikTok, Snapchat and Instagram [4][5]. So contact opportunities on Facebook specifically are limited, and a Meta-wide or cross-app estimate would be materially larger. Roughly 60% of online abuse also involves someone already known to the child [11], whom a stranger-contact model will not see at all.

Reading this responsibly

What would change the numbers

- **The base rate.** Move it from 0.02% to 0.20% and every 'real' figure moves roughly ten-fold. This single assumption dominates the estimate.
- **The platform.** Facebook alone understates Meta-wide risk; teens are mostly elsewhere. Treat this as 'Facebook as an example', not the whole picture.
- **Ranking quality.** Real detectors degrade on coded language, encrypted chat and new platforms; the illustrative AUROC is optimistic for a live deployment.
- **Age inference.** Guessing who is a minor carries a 4–7 year error; cases resting on shaky age estimates deserve lower confidence.
- **A snapshot.** These are steady-state figures; offenders adapt once detection exists, so any real number is time-limited.

The safety rules are not optional

- **People decide.** A score only orders the review queue. It never suspends, reports or names anyone by itself.
- **Most flags are false alarms.** The tables above make that explicit — which is why every flag gets a human, and why flags stay inside the safety team.
- **No public naming, ever.** A false accusation of this kind destroys an innocent life; watch-lists and exposure are prohibited.
- **Confirmed cases go to lawful bodies** — in the US the NCMEC CyberTipline; the IWF in the UK; EU trusted-flagger channels — initiated by a human finding, and logged.
- **It is auditable.** Every flag, decision and escalation is recorded, and fairness is checked across languages and groups.

> **Bottom line**
> For Sweden on Facebook, a reasonable base-case picture is: a few thousand genuine risk accounts exist; a review shortlist of similar size would surface several hundred of them with most flags being false alarms; and the tighter you set the dial, the more accurate — but narrower — the result. Every figure here is a planning estimate built on illustrative detection rates, to be replaced by measured values from a real evaluation before any reliance.

References

[1] Statistics Sweden (SCB), Population statistics 2025 (publ. 24 Feb 2026). scb.se

[2] DataReportal, Digital 2025: Sweden (Jan 2025). datareportal.com/reports/digital-2025-sweden

[3] NapoleonCat, Facebook users in Sweden, 2025–2026. napoleoncat.com

[4] Mediemyndigheten (Swedish Media Authority), Ungar och medier 2025. mediemyndigheten.se

[5] Internetstiftelsen, Svenskarna och internet 2025 (Facebook). svenskarnaochinternet.se

[6] NCMEC, CyberTipline 2024 data (publ. 2025). missingkids.org

[7] Thorn, What the 2024 NCMEC CyberTipline Report says about child safety (2025). thorn.org

[8] Meta Transparency Center, Community Standards Enforcement / NCMEC reporting, 2023–2024. transparency.meta.com

[9] Meta teen-safety enforcement: ~600,000 accounts removed for predatory behaviour, Aug 2023 (reported 2023–2025).

[10] Dunkelfeld / Troubled Desire (Charité Berlin): ~1% of men report paedophilic interest. troubled-desire.com

[11] WeProtect Global Alliance, Global Threat Assessment 2023 (60% of abuse involves someone known to the child). weprotect.org

[12] Fry et al., Online child sexual exploitation prevalence: systematic review and meta-analysis, Lancet Child Adolesc Health (2025).