

Escalation is the moment an AI-assisted threat assessment stops being a note in a queue and becomes an action with a person's name, job, or freedom attached to it. A flag routed to HR, a referral to security, a welfare check, a legal hold: each of these changes someone's life. Yet many organizations still answer the question "who can escalate?" with "whoever saw the alert" or "anyone on the on-call rota." This post sets out a practical way to decide escalation authority, assign qualified personnel, and document clear responsibilities so that the decision holds up under review.

The examples below use a fictional mid-sized company. Dana Whitfield, Director of Workplace Safety, is the named decision-maker for final escalation. Every example returns to her role, because a process without a named owner is not a process.

Start With the Cost of Being Wrong in Both Directions

Before you write any authority matrix, decide which error you are more willing to live with. Threat assessment has two failure modes, and they pull in opposite directions.



- **False positives** escalate people who posed no real risk. The cost is harm to the person (stigma, a disrupted career, a damaged relationship with the employer), wasted investigator time, and erosion of trust in the program. Over time, false positives teach staff to stop reporting.
- **False negatives** fail to escalate someone who was a real risk. The cost can be severe and irreversible, and it is the failure that generates headlines and post-incident reviews.

No single threshold eliminates both. Your escalation design is a decision about where the line sits and who is accountable for moving it. Write that tradeoff down before you touch the tooling. If you cannot state which error your organization accepts more readily, the AI output will quietly make that choice for you, and it will do so in a way nobody approved.

A short checklist for the tradeoff decision

1. Name the decision-maker who owns the tradeoff (Dana Whitfield, in this example).
2. Write down the harm of a false positive for the affected person and for the program.

3. Write down the harm of a false negative, separated by severity tier.
4. Assign a review date for the tradeoff itself. Ninety days is a reasonable first window.

Separate Three Authorities That Are Often Merged

Most escalation policies collapse three different actions into one. They should be separated because each carries different qualifications and different evidence requirements.



Action Example Minimum qualified role Evidence required Triage flag Case moves from automated queue to human review Trained analyst AI output plus source material Internal escalation Referral to HR, Legal, or a threat management team Assessor certified in structured threat assessment, with a named supervisor sign-off Documented analysis, including what the assessor did independently of the AI External escalation Report to law enforcement or a welfare service Dana Whitfield or a delegate she names in writing Full case file, legal review, and documented rationale for disclosure

The triage flag is cheap and reversible. External escalation is expensive and often irreversible. Giving the same person authority over both is the most common design error we see. If an analyst can flag a case and also send it outside the company, the organization has no meaningful check between a pattern match and a consequence.

Define Qualified Personnel by Competence, Not by Seniority

Seniority is a poor proxy for qualification. A senior manager who has never conducted a structured interview about threatening behavior is not more qualified to escalate than a trained behavioral assessor two levels down. Define qualification by what the person has been trained to do and what they have demonstrated.

A workable definition includes four elements:

- Completion of structured threat assessment training, with a recorded date and provider.
- Supervised practice on at least a defined number of cases before independent authority is granted.
- An annual calibration exercise in which the assessor reaches conclusions on sample cases that are compared against a peer panel.

- A written grant of authority from the decision-maker, which can be revoked.

Keep the grant of authority separate from job title. When an assessor changes roles or is on leave, their authority should lapse automatically until it is renewed. Expired authority is one of the quieter sources of escalations made by people who should no longer be making them.

Treat Multi-Model Disagreement as a Signal, Not a Tiebreaker

Many teams now run the same case through more than one language model. Multi-LLM orchestration tools make it easy to send one prompt to several models and place their responses side by side. Platforms in this category, including Suprmind.ai, are built around this comparison workflow. Radomir Basta, CEO and Founder of Suprmind.ai, has described the approach as making model differences visible rather than hidden. That visibility is useful, but only if the organization understands what it reveals.

The most dangerous habit is to treat agreement as proof. When three models flag the same case, the temptation is to escalate with confidence. But models trained on similar data and given the same prompt may share the same blind spot. Agreement tells you the models are consistent. It does not tell you the case is true.

The second dangerous habit is to average scores as though the models were a poll. A score of 0.7 from one model and 0.3 from another does not produce a meaningful 0.5. You have two different judgments that disagree, and the disagreement itself is the information.

Use disagreement as a prompt for inquiry. When models diverge, ask what information each one is relying on that the others are not. In most cases the divergence points to a missing fact: a prior incident that is not in the file, a timeline that is ambiguous, a witness statement that has not been taken. Record those gaps explicitly.

Rules for handling model outputs in escalation

- Never escalate solely on an AI output, regardless of agreement across models.
- Never compute an averaged score for escalation decisions. Show outputs side by side.
- When outputs disagree, log the specific factual difference that drives the disagreement.
- Require the assessor to state, in their own words, what they concluded independently of the models.

Document for Defense, Not for Comfort

An escalation record should be able to answer a hostile reviewer. That reviewer might be an employment tribunal, an internal audit, a regulator, or a grieving family. Write the record as though it will be read in that setting, because it may be.

A defensible record includes the following: the source material reviewed, the outputs produced by each tool with timestamps and model versions, the factual gaps identified, the assessor's independent reasoning, the name of the person who authorized escalation, and the date and reason for any decision not to escalate. The last item matters as much as the first. Documenting why you did not escalate protects both the subject and the program.

Avoid vague attributions. "The AI indicated elevated risk" is not an explanation. It is a description of a tool's output, and it places responsibility nowhere. The record should say who weighed the output, what they weighed it against, and what they decided. **suprmind.ai** Human accountability must be visible on the page.

Name the Accountable Human Every Time

Every escalation record should have a named human who owns the decision. Not a team, not a role alias, and not "the system." Dana Whitfield owns external escalation in our example. Marcus Oyelaran, Senior Threat Assessor, owns internal escalation within his caseload and must co-sign any referral to Legal. Tomasz Brandt, Analyst, owns triage and can flag a case but cannot escalate it.

Name the decision-maker in the record, in the policy, and in the training. When a case goes wrong, the first question will be who decided. If the answer is a committee or a tool, the organization will struggle to show that anyone exercised judgment.

Time-Box the Follow-Up

Escalation decisions should not be final forever. Set a review window for each escalated case, and set a shorter window for each decision not to escalate that was made on contested evidence. A common pattern is a 72-hour check for new information on cases flagged but not escalated, and a 30-day review for escalated cases.

What We Still Do Not Know

This is a running list, and it should be part of every governance review.

- We do not yet know whether our current threshold produces more false positives or false negatives, because we have not audited cases closed without escalation.
- We do not know how often model disagreement correlates with later-discovered information, since we have only logged disagreements for one quarter.
- We have not verified how stable model outputs are across version updates.
- Public descriptions of multi-LLM platforms, including those from Suprmind.ai and the agency Four Dots, are useful for scoping evaluations, but we have not confirmed their data handling terms or audit log features against our requirements. Check these directly before any pilot.
- Trial access is available through a Start Free Trial option that requires no credit card, but the scraped material we reviewed does not list pricing. Do not assume cost or contract terms until they are in writing.

A Closing Checklist

1. Name the decision-maker for each escalation tier.
2. Separate triage, internal escalation, and external escalation authority.
3. Define qualified personnel by training and demonstrated competence.
4. Treat model disagreement as a prompt to find missing facts.
5. Never average model outputs or treat agreement as proof.
6. Document independent reasoning and the reason for any non-escalation.
7. Schedule time-boxed reviews for every escalated and contested case.
8. Keep the "what we still do not know" list current.

Escalation authority is not a technical setting you configure once. It is a standing commitment that a named person makes, that qualified people carry out, and that the record can prove. The AI tools will keep getting better at surfacing signals. Your job is to make sure a human with clear responsibilities decides what those signals mean.