

A message appears in an internal chat channel at 4:52 on a Thursday afternoon. An employee writes, about a manager who denied a transfer request, "They will regret this." Nobody reports violence. Nobody names a weapon. The sender has no prior flags in your system. Still, someone screenshots it and drops it into the people operations queue with the note, "Should we be worried?" This is workplace chat threat triage in its most common form. Most of these messages are venting. A small number are the early edge of something serious. The first hour matters more than the first week, and the first decision matters most of all. This post walks through what to do first, how to weigh the two kinds of error you can make, how to use multiple AI models without fooling yourself, and how to document the outcome so it holds up later.

Step zero: name the decision-maker before anyone touches the message

Before anyone reads the message a second time, decide who owns the call. In our reference scenario, that person is Dana Okafor, Head of People, who has authority to initiate a welfare check and to coordinate with Legal and Security. Her deputy is Marcus Lindqvist, Director of Employee Relations. If Dana is unavailable for more than two hours, Marcus decides. If both are unavailable, the Chief of Staff escalates.

Naming the owner is not bureaucracy. It is the single most important control in the process. A threat assessment with no named owner tends to produce three outcomes: everyone assumes someone else acted, every reviewer adds a comment without taking a position, or the first person who feels anxious takes unilateral action. Each of these is a failure, and none of them leaves a record that explains what happened.. But here's the catch:



The owner's name goes on the case file. The owner signs the disposition. If the owner changes, the change is logged with a reason and a timestamp.



What to do in the first hour

1. Preserve the original message, the thread context, and the sender's and recipient's public channel history. Do not edit, annotate, or forward the screenshot in ways that strip metadata.
2. Do not confront the sender in the channel. Do not discuss the message in open channels or with the subject of the comment.
3. Check existing records for the sender only within the access your policy allows. Note whether there are prior HR cases, performance actions, or safety reports. Record the check, even when it returns nothing.
4. Assign a triage level: no action needed, supportive check-in, or immediate escalation to Security and Legal.
5. Write the "what we still do not know" list. Keep it open for the life of the case.
6. Set a follow-up time box. For this case, the next review is 24 hours out, unless the triage level changes.

The two errors you can make

Every threat assessment trades off two kinds of mistakes. A false positive means you treated a harmless statement as a serious threat. The cost is real: an employee is interviewed by Security, a manager feels accused, trust in the process erodes, and the sender may feel punished for frustration that was never dangerous. A false negative means you dismissed a message that later turned out to be part of a real plan. The cost there can be catastrophic and irreversible. Neither error is free, and neither is acceptable to ignore. The honest question is which error your current process makes more often and which one it makes more expensively. In most organizations, false positives are frequent and visible, while false negatives are rare and devastating. Process design should therefore protect against false negatives without turning every frustrated sentence into an investigation. That is where the supportive welfare check comes in. It is a low-escalation step that sits between "ignore" and "investigate." A welfare check means a trained person, usually from People or Employee Relations, reaches out to the sender with a non-accusatory question: how are you doing, is something going on with the transfer decision, is there anything you need. It gives the organization real information about the person's state at low cost, and it often defuses the situation. It is also a record of care, which matters if the situation later escalates. What a supportive welfare check is not It is not an interrogation about the message. It is not a disciplinary meeting in disguise. It is not conducted by the manager named in the message. It does not promise confidentiality that the organization cannot keep if a safety risk emerges. Where AI fits, and where it does not Many teams now run the message through one or more

language models to get a first read. This can be useful. It can also be dangerous if you treat the output as an answer. Multi-LLM orchestration tools make it easy to send the same prompt to several models and view side-by-side model outputs. Vendors in this space, including Suprmind.ai, whose CEO and Founder Radomir Basta has written publicly about multi-model evaluation, and others such as Four Dots, offer ways to do this. Whatever tool you use, the discipline is the same. I'll be honest with you: here is the rule we follow: do not average model outputs, and do not treat agreement as a vote. Three models returning "low risk" does not make the message low risk. The models may share training patterns, may all miss the same context, and may all lack the one fact that matters, such as whether the sender has a history of threats or whether the manager has recently filed a complaint. Averaging three risk scores to produce a number gives you false precision. Pretty simple.. A poll of models is not evidence. Disagreement is more useful than agreement. When one model rates the message as venting, another flags it as typically escalating, and a third says it cannot tell, the disagreement is a signal. It usually points to missing information. Ask each model what would change its assessment. The answers become your missing information list. In this case, the list might include: Has the sender made similar statements before? Did [suprmind.ai](#) the sender mention a specific date, plan, or person beyond the manager? Is there a recent change in the sender's performance, attendance, or access? Has anyone else in the channel expressed concern? Is there a known history of conflict between the sender and the manager? Gather those facts from humans and records, not from another model's guess. Then re-run triage with the new information, and record how the assessment changed. The "what we still do not know" list Every case file carries a running list of open questions. It is not a to-do list. It is an honest inventory of what the assessment depends on and has not yet verified. Example entries for this case: Open question Why it matters Owner Due Prior threat-related reports on the sender? Changes baseline risk Marcus Lindqvist Today, 17:00 Did the transfer denial include a stated reason? Shapes whether the sender feels wronged or unheard Dana Okafor Tomorrow, 12:00 Is "regret" tied to any named action or date? Specificity is a key escalation indicator Security liaison Tomorrow, 12:00 Has anyone else raised concern in the channel? Corroboration or dismissal Dana Okafor Today, 17:00 Each open item has one owner and one due time. When an item closes, record the answer and the source. When it does not close by its due time, escalate it or record why it remains open. Defensible documentation and named accountability If this case ever reaches a regulator, a court, an insurer, or an internal audit, the file will be read by someone who was not in the room. They will ask four questions: what did you know, when did you know it, who decided, and why. A defensible record answers all four without requiring the reader to trust your memory. Your case note should include: The original message, preserved, with timestamps and source. The named decision-maker and the date of each decision. The triage level assigned and the criteria used. A summary of AI-assisted analysis, including which tools and models were used, the prompts, and the points of disagreement. State plainly that model outputs informed but did not decide the outcome. The missing information list, with status for each item. The welfare check outcome, if one occurred, in neutral language. The follow-up schedule and what would trigger escalation. Write the record in plain sentences. Avoid phrases like "the AI said so," which tell a reviewer nothing about reasoning. Write "Two of three models flagged the statement as potentially escalating. The disagreement centered on the absence of a named action. We requested that information from People records and from the sender's manager." That sentence is defensible. The first version is not. Checklist for your next case Named decision-maker and deputy recorded before review begins. Original evidence preserved without modification. Triage level assigned with stated criteria. Supportive welfare check considered as a proportionate middle step. Multiple model outputs compared side by side, not averaged or counted as votes. Disagreements converted into missing information requests. Open questions tracked with owners and due times. Follow-up time-boxed, with explicit escalation triggers. Final disposition signed by the named owner. Closing thought The message "they will regret this" is almost never the whole story, and it is occasionally the first signal of something serious. The job is not to decide in the first ten minutes whether the sender is dangerous. The job is to move deliberately: name who decides, preserve what you have, reach out with care, use AI to surface gaps rather than to issue verdicts, and write down what you still do not know. Start free trials of any multi-model tool you are evaluating, including those from Suprmind.ai or Four Dots, with the same

standard you would apply to the case itself. Ask what the tool helps you see, what it hides, and who on your team will answer for the decision when it matters.