CHAPTER 3

Identifying and Responding to Online Harm and Conflict

Moderators (mods) don't need to know every law or policy word-for-word, but you do need a simple framework for what to do when harmful content or escalating conflict shows up.

3 key questions before acting

1. Could it harm someone?

- Might a post cause serious emotional distress to someone in the group?
- Does it include physical threats or anything that could lead to real-world harm?

2. Does it break the rules?

- Does it go against platform community standards/or the HDCA?
- Is it breaking your group's kaupapa or rules?

3. Could things get worse?

- Are arguments or pile-ons already forming in the comments?
- Is this the kind of post that could snowball if left alone without admin intervention?

Click to watch Spotting Harm & Calming the Storm

Follow this cycle:

Identify → Respond → Prevent Recurrence

1 Identify

Spot harmful, illegal, or high-risk content quickly. Look out for anything that could cause serious emotional distress, spread false allegations, or break either your group's rules, the platform's community standards, or the HDCA.

Keep watch for conflict risk factors:

- Pile-ons where one member is ganged up on as a form of harassment or abuse
- Threads drifting off-topic in a way that reduces constructive engagement and increases harmful interactions or escalating online fights
- "Just asking questions" or other dog-whistling tactics
- Repeated snark, sarcasm, or baiting that hinders constructive dialogue and the quality of engagement
- Situations that present serious or imminent risk of harm or violence

2 Respond

Take clear, consistent action to warn, de-escalate, hide or remove. This might mean removing harmful content, issuing a warning, muting a member, or escalating to Netsafe or Police if the situation is unlawful, criminal or serious.

Manage conflict before harm escalates:

- Step in early with a tone-setting comment
- Use comment slow mode to cool things down
- Freeze a post (turn off comments) if kōrero gets hostile and counter productive
- Remove individuals, not ideas ("This post is being paused due to tone, not your opinion"). Moderation actions focus on the manner of expression not the viewpoint.
- Use calm, neutral language: "We've paused this thread while we review comments. Please stay respectful and follow our group's rules."

3 Prevent Recurrence

Learn from each incident. Log it in a mod record, tighten your group rules if needed, and use moderation tools to stop repeat problems before they start.

Consider additional conflict-prevention practices:

- Post "tone reset" statements after heated threads
- Add reminders or updates to your rules periodically
- Use a monthly group reset post to re-centre the kaupapa
- Thank members who help de-escalate, modelling respectful behaviour helps set the culture

This simple cycle protects your group, builds trust with members, and shows you've taken reasonable steps if your moderation decisions are ever questioned later.

Conflict in online spaces isn't always bad, but how it's handled can make or break your group.

Type	Description	Examples
Genuine disagreement	Members have different life views or values	Politics, religion, culture, identity
Tone escalations	Good-faith discussion turns snarky, sharp, or defensive	Accusations, sarcasm, "pile-ons"
Bad faith baiting	Someone posts with intent to provoke or troll	"Just asking questions," dog-whistling
Hijacked threads	A normal post gets overrun by off-topic fights	Vaccine memes on a housing thread
Identity targeting	Arguments become discriminatory	"You people always…", coded slurs
Post-incident backlash	A harmful post or removal causes member unrest	"Why was I banned?" "Mods are silencing us!"

If you have laid a strong foundation through consistent and balanced moderation efforts, community members will often help you uphold the tone and values of the group by responding to agitators with reminders of rules, reporting harmful content to mods, and calling out bad behaviour before it escalates.

> ***"There's a really solid group who'll step in and say, 'That's not what we do here.' They'll gently remind others that naming and shaming isn't part of this space and by doing so, it resets the tone"***
>
> **- anon, NZ mod**

Harm types:

- Scams & frauds
- False allegations (Mis/disinformation)
- Graphic/violent imagery
- Bullying and harassment
- Deepfakes and synthetic content
- Gang-related or extremist content and intimidation
- Doxxing or exposing private information
- Unverified or misleading theories
- Identity-based abuse
- Encouragement or details of suicide or self-harm
- Discriminatory gendered abuse or online extremism rhetoric

What to watch for (online risk factors)

- Sudden spikes in toxic or polarised comment threads
- Floods of AI-written or copy-paste information from fringe or unverified sources
- Links to unverifiable or anonymous websites and images
- Conspiratorial frames (e.g. "they don't want you to know this")
- Disclosures of private/sensitive information
- Elevated or serious emotional and/or hostile reactions to any mentions of Aotearoa or te reo Māori words
- Displays of gang insignia, coded threats
- Extremist recruitment materials
- Elevated or serious emotional and/or hostile posts targeting or diminishing identities (e.g. Gender, race, religion)
- Hijacking of posts to spread unrelated controversial content

> ***"That kind of behaviour creates a lot of negativity. When it happens, we step in, posts are taken down, and we remind people to stay respectful. There have been times when the whole moderation team has had to get involved."***
>
> **- anon, NZ mod**

EXAMPLES YOU MAY HAVE COME ACROSS IN NEW ZEALAND

Scenario	Example	Online Risk Factors
Public health misinformation	Pandemics	Viral meme links to false information, fake statistics, falsifying circumstances of deaths
Natural disaster events	Cyclone recovery donation scams	Emotional posts, “share this fast” claims, charity impersonation
Racism	Anti te reo Māori comments, posts blaming immigrants for unemployment/ housing shortages	Identity-based attacks, blame narratives, attacks on culture or language
High emotion protests	Global conflicts, occupations, animal, fauna and pest control	Hashtag storms, cross-posted rage posts designed to cause online and/ or offline harm
Terror attack	Domestic and international terrorist events, hate speech and trolling	Extremist images, incitement memes
Election period	General elections party-based fights in comments detracting from constructive political discussion	Meme flooding, campaign disinfo, candidates being targeted that can risk physical attacks or mislead audiences and disrupt informed debate.
Gang activity	Posts glorifying gangs or threatening rivals	Symbols, videos of weapons, threats to life
Extremist recruitment	certain online content urging people to “protect NZ’s identity” and/or join “patriot” groups, linking to alternative messaging spaces for recruitment	Exclusionary framing, links to extremist messaging platforms, calls for offline action, coordinated sharing of extremist viewpoints for the purpose of radicalisation
Animal groups	Comment wars over diet, breeding ethics, showing animals, rescue vs breeder care	Threads quickly polarise. Emotional or accusatory tone. Accusations of neglect/abuse.
Gendered abuse	Anti-feminist memes, promotion of tech facilitated abuse and victim blaming	"Red pill" content, targeted comment chains, coercive control instructions
Anti-rainbow rhetoric	Posts attacking Pride celebrations	Targeted harassment, deliberate misgendering
Misogyny/online discussion of masculinity	Incel-related slurs, minimising women’s issues	External links to discriminatory forums, dog whistles
False or misleading information	5G towers and overstated health risks, cloud seeding etc.	Hashtags, fringe YouTube links, anti-corporate sentiment
Community or public figure death (suspected suicide)	A member’s post speculates about a local person or celebrity death, saying they “committed suicide” “by the tracks.”	Speculation about cause, location and method, glamorising tone, repeated sharing, copycat/contagion risk

Approaches you could take

Depending on the context and community, you'll learn which approach works best. When tricky or potentially harmful content appears, the goal isn't to silence discussion — it's to keep kōrero safe, factual, and respectful. You might choose to pause or slow comments, post a holding statement, or share a verified source to correct misinformation. If tension rises, lock or limit threads, remind members of the group kaupapa, or post temporary tone guidelines to reset the conversation. When harm occurs, remove or hide content, log incidents, and offer support to those affected. For ongoing or serious issues, you may need to escalate to Netsafe or NZ Police.

Handled calmly and transparently, good moderation protects both free expression and community wellbeing.

When to escalate

If an online situation objectively appears unsafe and harm is occurring, escalate by doing the following:

Escalation Path	Use When...
Another mod	You're unsure, triggered, or overwhelmed
Lead/admin mod	Action may impact group direction or trust
Netsafe	Helping people to identify and deal with HDCA matters, including harassment, threats, doxxing and more.
Police	Threats of violence, child safety risk, suicide threats, extremism concerns
Community orgs	As relevant to the situation - i.e suicide helpline, health helpline, Outline, Crisis Line 1737, etc

Set up a mod group chat for real-time coordination

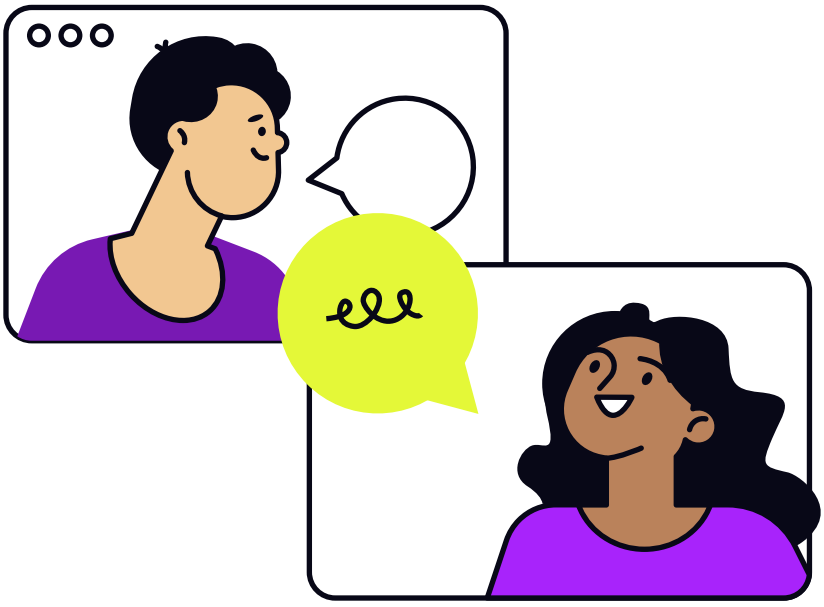
"Setting up a Facebook Messenger just for the moderators in our group worked really well. This way we could comfortably share our experiences, offer advice to each other, as well as explain our thinking behind the decisions we made so that we were all on the same page." **- anon, NZ mod**

Discussion prompts for your mod team:

- Do we agree on our group's threshold for harm?
- How do we handle members who unintentionally spread misinformation?
- Are our members aware of our no-doxxing policy?
- Do we have a pinned post or FAQ to redirect members?

TIPS:
Trust patterns, not just words

- Look for who is being targeted or silenced
- Don't require hate to be explicit, harm often isn't
- Back each other up when calling these out

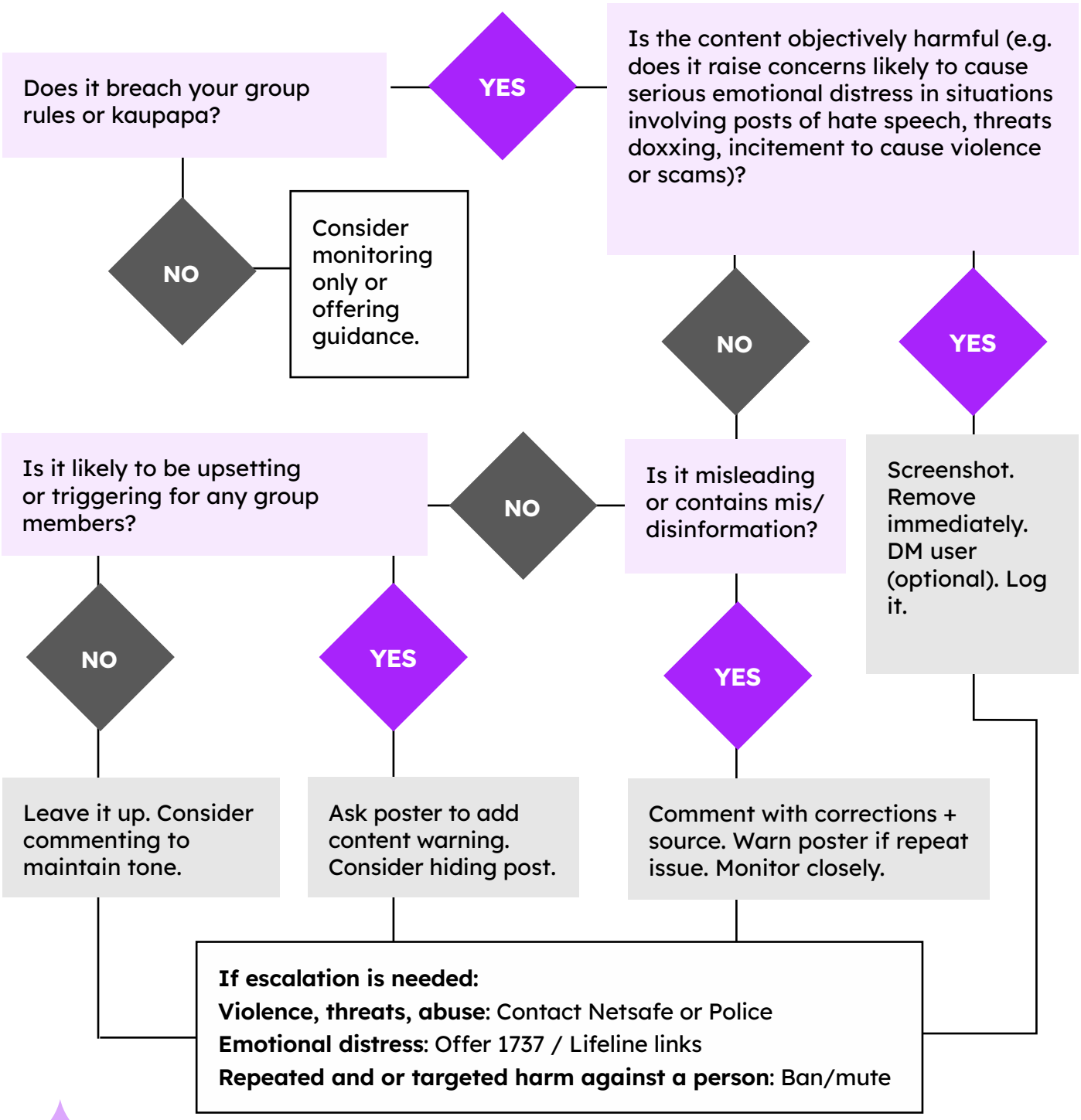


Decision trees for moderation

Use these yes/no flowcharts to guide decisions when tension or harm arises. They help take the emotional weight out of decision-making by giving you a clear path forward.

Content risk decision tree

Use this when reviewing a questionable or reported post

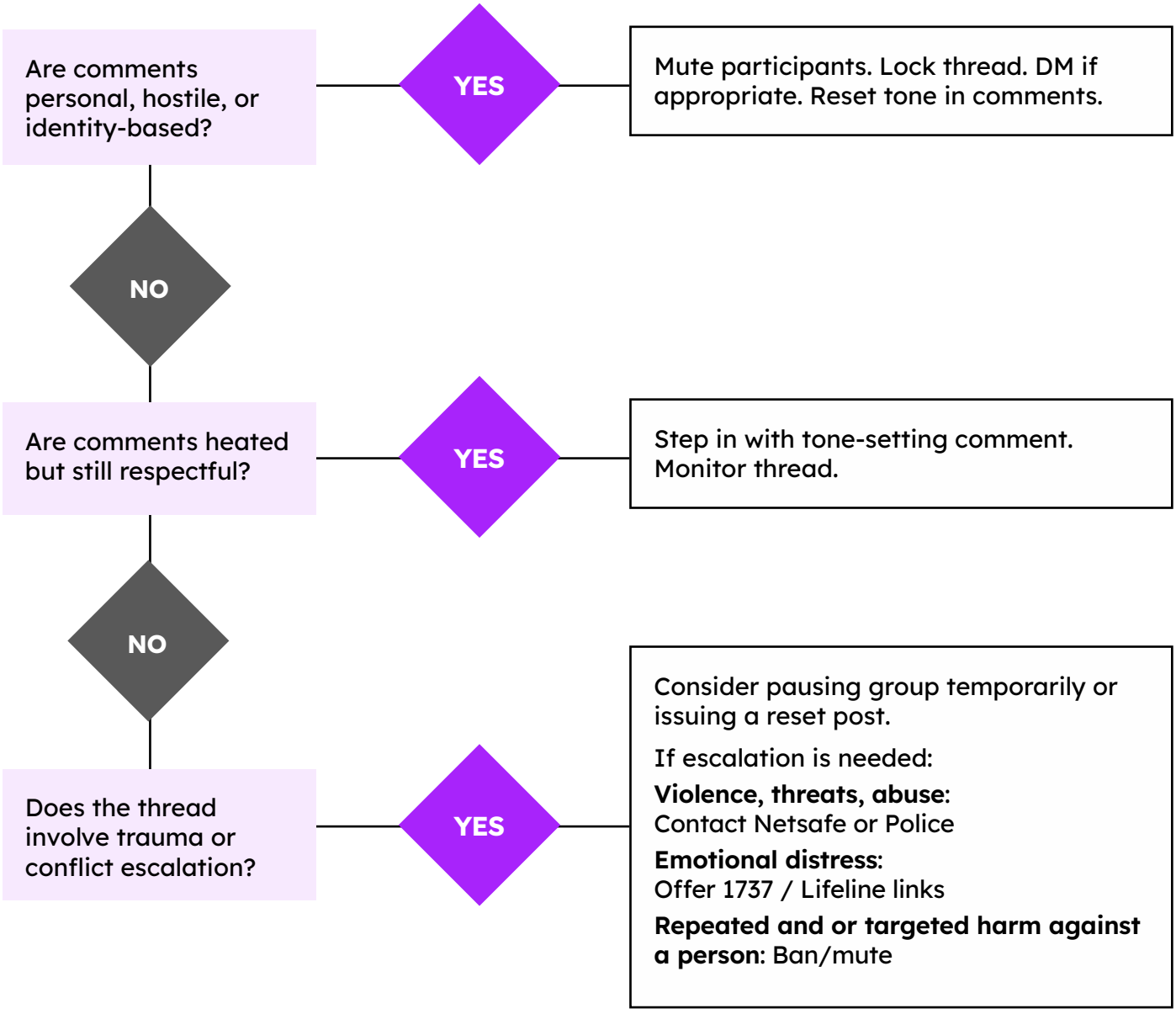


STILL UNSURE?

Add a “Holding Comment” and DM the user privately. Flag for discussion in mod chat.

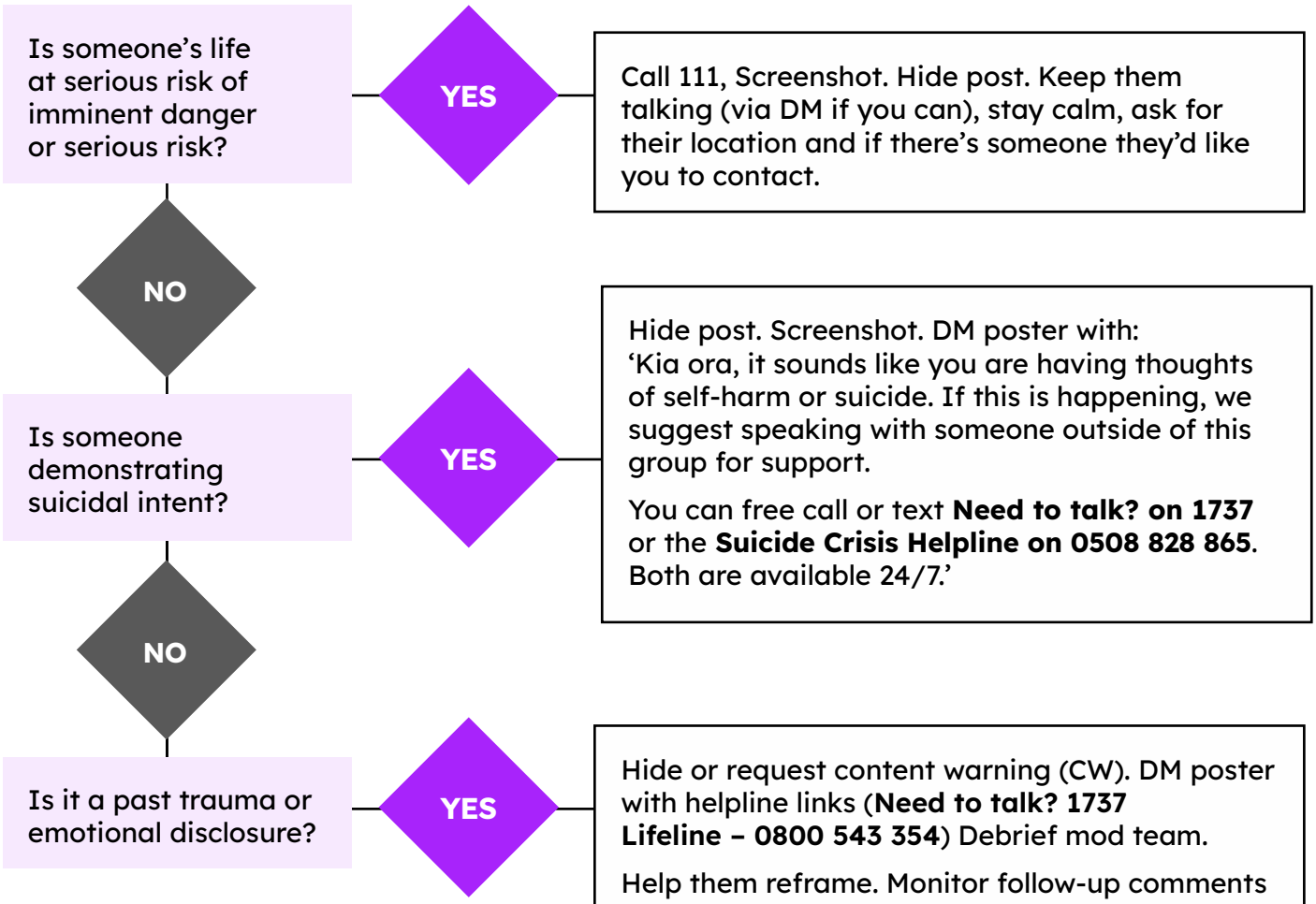
Comment fight decision tree

Use this when tension is rising in the comment section



Suicide or self-harm disclosure tree

Use this when a post shares experiences of self-harm or indicates suicidal intent. Treat all threats of self-harm seriously. If the person is in immediate danger, call 111.



TIP:
Moderate closely if there's activity suggesting a suspected suicide in the community. Safe language and images matter. It is a breach of the law (s.71 of Coroners Act 2006) to name the method, the location if it indicates method, and calling it a suicide. It's a suspected suicide until Coroner confirms otherwise.

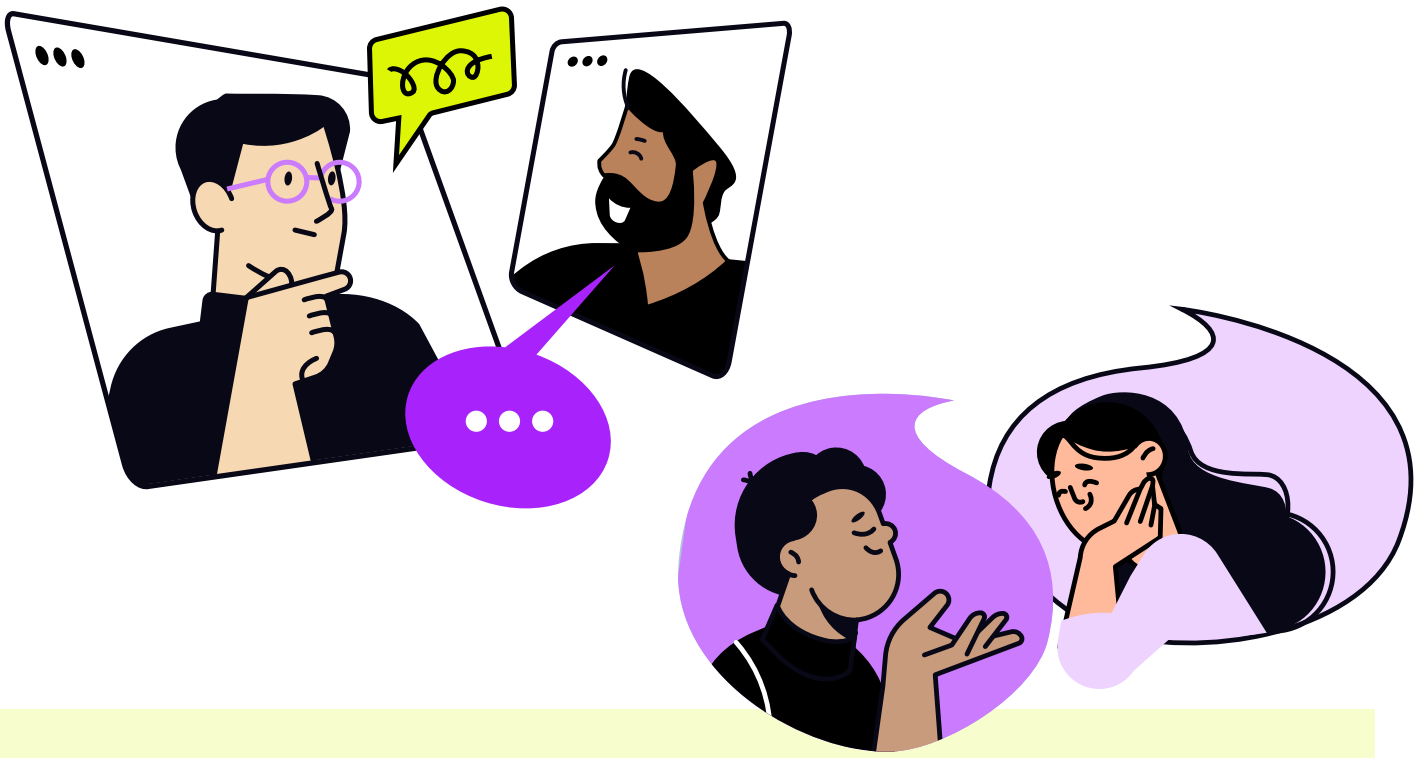
Moderator language library
These suggested phrases are tailored for Aotearoa New Zealand-based groups and prioritise kindness, clarity, and tikanga.

PUBLIC COMMENT TEMPLATES (VISIBLE IN THREADS)

Context	What to Say
Misinformation post	"Thanks for sharing this. Just a heads-up: the info isn't quite accurate - here's a verified source from (insert official source). Let's stick to facts here to keep everyone safe."
Conflict/flame war	"We're going to pause this kōrero for now while things cool down. Please come back with care and respect for our community when you're ready."
Graphic or triggering post	"Kia ora, please consider adding a content warning at the top of posts with heavy topics (e.g. Trauma, domestic violence). That helps others protect their wellbeing." However mods should immediately remove anything harmful along these lines.
Privacy breach or doxxing	"We've removed this post to protect someone's personal info. Please don't share names, messages, or screenshots without consent."
Hate crimes or abuse	"This post goes against our group kaupapa. We stand for respectful and inclusive kōrero, any kind of hate crimes isn't tolerated here."
General reset	"Let's reset the tone here. We're all here for different reasons, but we share a space where kindness and honesty can coexist. Ngā mihi for your patience."

PRIVATE MESSAGE TEMPLATES (DIRECT MESSAGES)

Situation	What to Say
Warning someone calmly	"Kia ora, just letting you know your recent comment went against our rule about respectful kōrero. We've taken it down. Feel free to repost if you want to reword it. Ngā mihi."
Supporting a distressed member	"Thanks for sharing something so personal. We've hidden the post for now just to protect your privacy and the group's wellbeing. If you'd like to repost with a content warning, that's totally okay. If you need help, 1737 is free to text or call any time."
On blocking or banning	"We've made the decision to remove you from the group due to repeated harm. Our kaupapa is grounded in harm minimisation, safety and care, this group may not be the right fit. All the best."



TIPS:

Always log major moderation decisions in your internal records.

Managing public backlash

Sometimes moderating fairly still gets your flak. Don't take it personally, use structure, systems and communication to help you manage tricky situations.

Steps:

1. **Acknowledge members' confusion** ("We hear there's concern about recent removals…")
2. **Restate kaupapa clearly** - rules, reasons, purpose
3. **Use a "holding post"** - a pinned explanation can calm chaos
4. **Use consistent language** as a team
5. **Mute or remove repeat agitators quietly** and without drama

MODERATOR SCRIPTS (QUICK USE)

Situation	Script
Off-topic rant	"This post is getting off-topic. Please bring it back to the group kaupapa or we'll need to close comments."
Low-level trolling	"This feels more like provocation than discussion. Tone matters."
Defensive reply to mod	"We're not here to debate moderation in-thread. Please DM us."
Subtle racism/ dogwhistle	"This language isn't acceptable here, even if it's implied. We prioritise safety over semantics."
Banned user's friend complains	"We're not discussing another member's situation. Thanks for understanding."

Identifying and Responding to Online Harm and Conflict

Live streams: handle with caution

Live streams can help connect your community, but they also come with serious risks. Harmful or illegal content can appear in real time, and once it's broadcast, it's hard to undo the damage. Moderators should take extra care before allowing or promoting live streams in any online community.

Best practice for live streams:

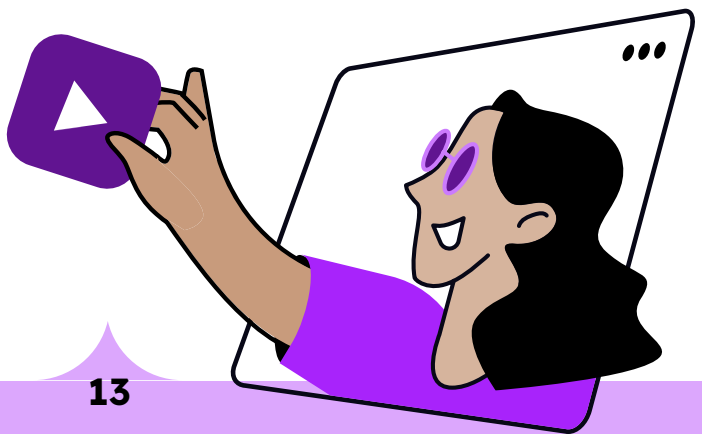
Turn off live streaming unless your community specifically needs it (e.g. planned Q&As, verified organisational broadcasts).

If it is a core part of your community approach to content, then:

Pre-approve who can go live and require mods to be present or monitoring during the stream.

Establish clear rules that any live content must:

- Follow community kaupapa and group rules.
- Not include minors without guardian consent.
- Avoid showing distressing or violent material.
- Respect privacy, no filming private property or individuals without permission.



If harm occurs live:

- End or suspend the live stream immediately by clicking the three dots ••• at the top of the post then "Remove post"
- If you can't remove it mid-stream:
- Mute the member, disable commenting
- Screenshot only the post title or user ID (not the stream)
- Report the user and escalate to Police or Netsafe if necessary
- Log the incident and notify the rest of the mod team

TIPS:
Treat live streams like public events, plan, moderate, and debrief them. Never assume "it'll be fine."

Monthly group reset post

Sometimes when things get a little heated or the group has experienced multiple incidents with conflict, it helps to remind everyone about the core kaupapa and rules of the group. Consider when you might need to add a 'reset post' and pin it for easy reference.

"Kia ora everyone, just a quick tone reset for our space. This group exists to uplift, not harm. Please revisit the group kaupapa and rules if it's been a while. We moderate with aroha, but also with boundaries. Thanks for helping keep this group a safe and inclusive space for all."