A GUIDE TO ONLINE COMMUNITY MODERATION

APPENDIX

ACTION TOOLKIT – QUICK DECISIONS UNDER PRESSURE

Follow the 3 steps:

Identify → Respond → Prevent Recurrence

1 Identify

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?

HELPFUL MODERATION HABITS

- Act swiftly before things spiral.
- Stay consistent across moderators.
- Keep a record (screenshots with names, dates, URLs).
- Encourage positive engagement.
- When in doubt, check HDCA/ Platform Community Standards.

2 Respond

Risk	What it looks like	Action
🟢 **Low**	Off-topic, small disagreement, not harmful	Monitor, remind gently
🟡 **Medium**	May offend, rule-bending, repeat issue	Remove content, send kind warning, consider mute
🔴 **High**	Clearly harmful, abusive, threats, scams, criminal	Remove immediately, report to platform, remove user, escalate if needed

Escalation path	Use When...
Another mod	You're unsure, triggered, or overwhelmed
Lead/admin mod	Action may impact group direction or trust
Netsafe	HDCA, including harassment, threats, doxxing and scam advice
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 etc.

3 Prevent Reoccurence

Other things to keep in mind

- **First offence vs. repeat offender** – respond proportionately.
- **Reactions of others** – is harm spreading or creating fear?
- **Tone/intent** – mistake vs. deliberate targeting.
- **Offline consequences** – could this lead to real-world harm?

Balancing expression with safety

- Groups have a Kaupapa, set tone and boundaries.
- Allow diverse views but draw the line at harm or identity-based attacks.
- Be transparent about rules.
- Encourage healthy kōrero, not harmful drama.

PRIVACY TIPS

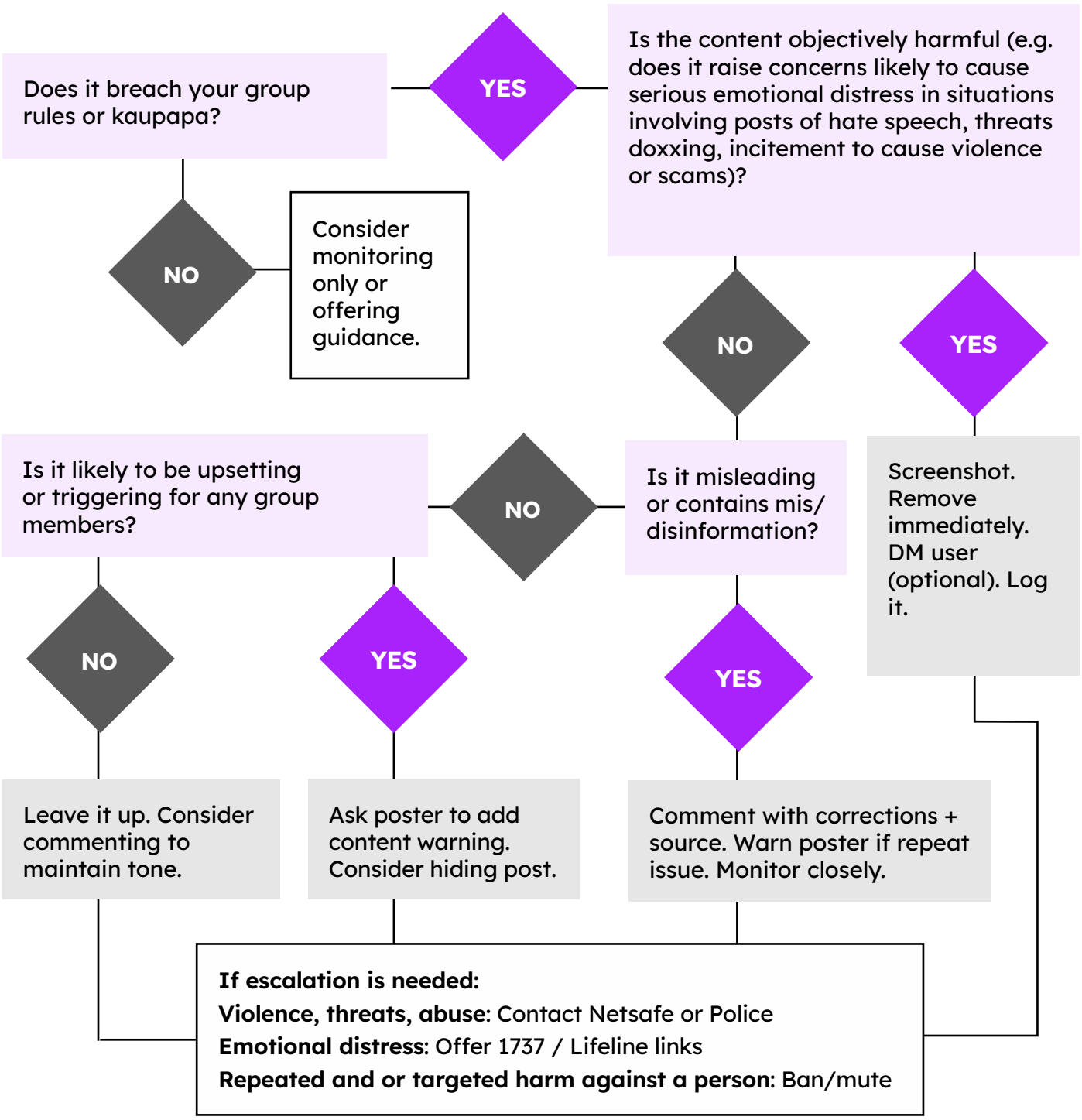
- Keep reports and warnings confidential.
- Store screenshots securely and only share within the mod team.
- Never reveal who flagged a post.
- Use private mod channels for team discussion.
- Ensure members protect their own privacy.

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

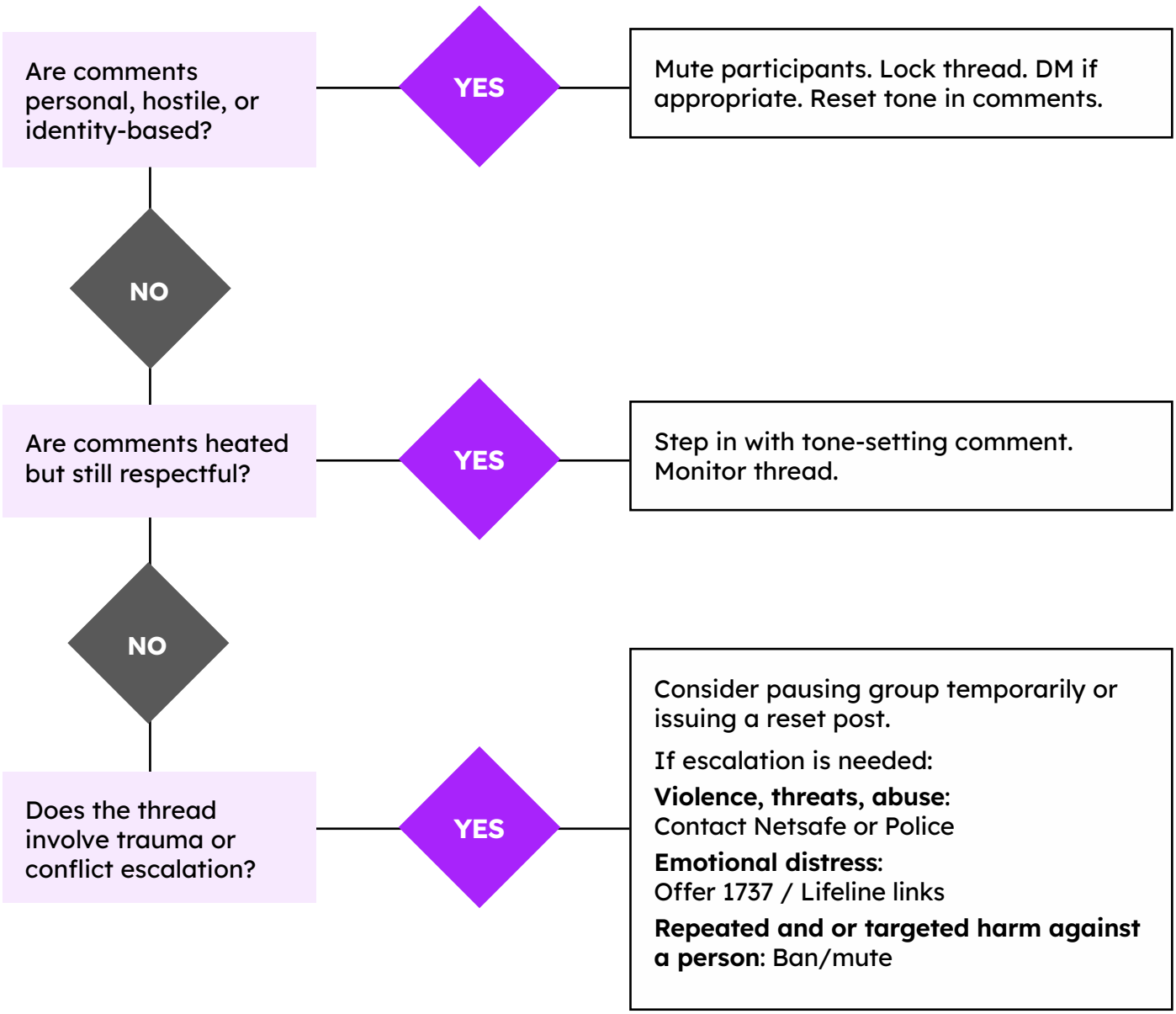


STILL UNSURE?

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

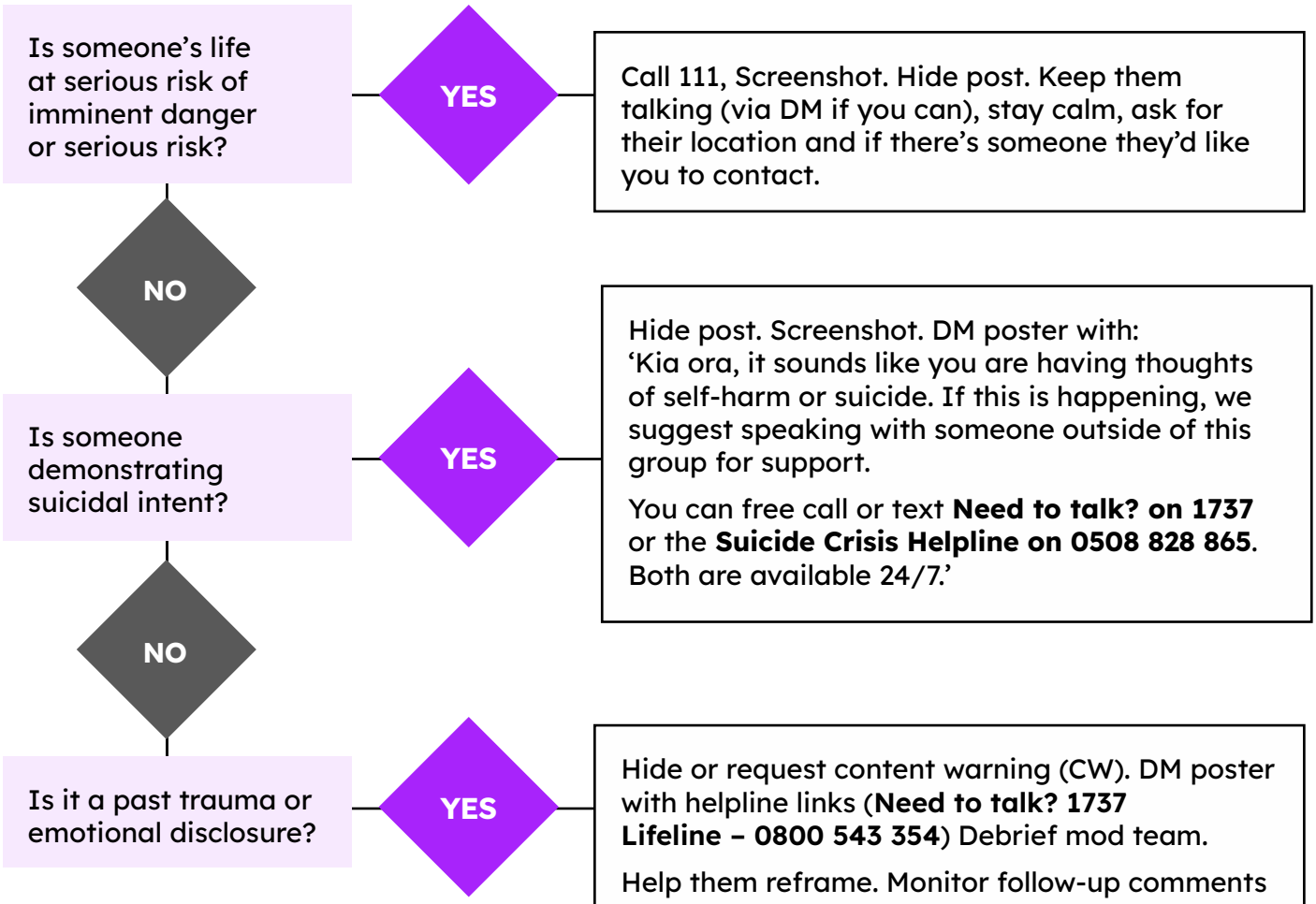
Comment fight decision tree

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.

AI FOR COMMUNITY MODERATORS

For time-poor moderators and those filling the role as a volunteer, AI tools can help lighten the workload and respond faster. Built-in AI enabled moderation tools are built into most platforms now and help mods set up a first line of defence against online harm. Other Generative AI tools can also be a useful helper for community moderators, but AI doesn't understand your group's kaupapa or tone unless you tell it, and it can make mistakes.

What you can use it for:

- Set up keyword filters, member screening and other back-end systems
- Drafting messages & holding statements – calm, neutral, supportive
- Summarising long threads – get the gist quickly without wading through
- Translating tone (for example, from too harsh to gentle, or too formal to friendly)
- Brainstorming tone-setting responses – when you're stuck for words
- Training support – practice scenarios, inclusive phrasing ideas

AI should help you hold the tone and save time, not replace your judgement.

Getting better results

AI can only work with what you give it. Vague prompts give vague results. The more context and instruction you include, the better the output will match your group's tone, kaupapa, and needs.

FOR EXAMPLE:

Poor prompt:
"Write a group post about respect."

Better prompt:
"Write a short, friendly post reminding members in a private Aotearoa parenting group to keep kōrero respectful during heated debates. Include a gentle reminder about our group rules and use plain New Zealand English."

"I'm a moderator of a private mental health support group in Aotearoa. Write a short message asking members to stop arguing and refocus on respectful kōrero. Keep it kind, calm, and mana-enhancing."

TIPS:

- Include context (who you are, what kind of group it is)
- Specify tone (calm, inclusive, mana-enhancing)
- Tell it how long or what format you need (two sentences, list, announcement)
- Mention what not to do (for example, "avoid emojis," "keep it neutral")

Iterating for better results

AI improves when you give feedback. Try adding:
“Make it sound more natural.”
“Use lighter language.”
“Shorten to two sentences.”
“Rewrite with a te reo Māori greeting and New Zealand spelling.”
“Keep it professional but friendly.”

Cultural and contextual awareness

AI tools are mostly trained on global data, not Aotearoa communities, so guide them to localised language and context.

Safe AI use

- Don’t share names, screenshots, or private details in your prompts Instead: paraphrase (“a member made a racist comment” rather than pasting their post).
- Don’t copy-paste confidential moderation discussions.
- Don’t rely on AI to decide what is harmful or illegal, escalate to Netsafe or Police if you’re unsure.
- Use trusted, reputable AI platforms.
- Avoid random plug-ins or bots that may store data unsafely.

Final check before posting

Before you post or send an AI-drafted message, check:

- Does it reflect your group’s kaupapa?
- Does it match your group’s tone?
- Is it accurate and neutral?
- Have you checked any links or sources of information for accuracy?
- Would you stand by it if someone asked how it was written?

AI can be a helpful co-pilot for moderators, not a replacement for empathy, cultural care, or accountability. Used wisely, it can help you focus on what matters most: guiding kōrero that keeps your community safe, balanced, and respectful. Always check, edit, and own what you post. You’re the human voice of your community.

For more in-depth information on safe and responsible AI use, see **netsafe.org.nz/artificial-intelligence**

GLOSSARY OF TERMS

1737 / Need to Talk?
New Zealand's free 24/7 mental-health support line. Call or text 1737 to speak with a trained counsellor.

Accessibility (Digital)
Designing online content so it's usable by people with disabilities — for example, alt text, captions, and clear layout.

Admin Assist (Groups)
A Facebook Group automation tool that pre-approves, declines, or flags posts and members based on rules you set.

AI (Artificial Intelligence)
Technology that performs human-like tasks such as filtering spam or summarising comments; used behind many Meta moderation tools.

Approved Agency
Under the Harmful Digital Communications Act 2015, Netsafe is the Approved Agency responsible for receiving complaints, assisting victims, and working with online content hosts to resolve harm.

Burnout
Emotional and mental exhaustion caused by prolonged stress; common among moderators exposed to harmful or high-volume content.

Communication Principles
Ten behavioural principles under the HDCA that outline what counts as harmful digital communication (e.g. no harassment, no false allegations, no incitement).

Community Standards
Platform rules on acceptable behaviour. Breaches can lead to post removal or account restrictions even if NZ law isn't broken.

Content Warning (CW)
A note at the start of a post alerting readers to sensitive or distressing content (e.g. violence, trauma, abuse). Sometimes know as Trigger Warning (TW).

Coroners Act 2006, Section 71 (Reporting Suicide)
Restricts publication of a suspected suicide's method or details until the Chief Coroner releases findings.

Crisis Terms
Words or phrases that signal distress or imminent risk and should prompt escalation.

Cultural Safety
Ensuring online spaces respect and protect the identities and wellbeing of people from all cultural backgrounds.

De-escalation
Moderator steps that reduce tension, using calm language, slowing comments, or temporarily pausing discussion.

Deepfakes
AI-generated or altered video, images, or audio that convincingly depict people doing or saying things they never did.

Defamation
Publishing false information that harms someone's reputation; truth and honest opinion are defences.

DIA – Department of Internal Affairs
NZ government agency responsible for addressing objectionable or illegal online content through its Digital Safety Group.

Digital Communication
Any form of online or electronic messaging, including emails, social media posts, comments, images, and videos.

Dog-whistle
Coded language that seems harmless publicly but signals bias or hostility to a targeted group.

Doxxing
Publishing or sharing someone's private or identifying information (address, workplace, phone number etc.) without consent.

Escalation
Referring a serious or unlawful issue (e.g. threat, harassment, exploitation) to Netsafe, Police, the DIA, or another agency.

Featured Post
A Facebook post pinned to the top of a group feed to highlight important or time-sensitive information.

FVPC Act (Films, Videos and Publications Classification Act 1993)
NZ law defining and regulating "objectionable" content such as child sexual exploitation, extreme violence, or terrorism material.

Guides (Groups)
A Facebook feature allowing key posts to be grouped into themes (Welcome, Safety, FAQs) for easy reference.

Harmful Digital Communications Act 2015 (HDCA)
NZ law to deter, prevent, and mitigate online harm and provide quick redress for victims of digital communications causing serious emotional distress.

Holding Post
A temporary pinned notice used during high tension to pause discussion and explain next steps.

Holding Statement
A neutral message to pause or de-escalate conflict (e.g. "We're reviewing this post, please pause kōrero for now.").

Kaupapa
A Māori term for purpose or foundational principle. In moderation, it anchors a group's values and decisions.

Kaupapa Māori
Māori-led frameworks and values that guide how communities are built and moderated.

Keyword Alerts (Groups)
Facebook tool that notifies moderators when chosen words or phrases appear in posts or comments.

Kōrero
Māori for "conversation" or "discussion." Used to describe respectful dialogue within online communities.

Live Stream
A real-time video broadcast. Moderators should monitor streams closely to prevent harmful or illegal content.

Mana
Authority, dignity, and prestige. Moderators maintain mana by acting fairly and respectfully.

Manaakitanga
A Māori value meaning care, hospitality, and respect — hosting others kindly and safeguarding community wellbeing.

Misinformation / Disinformation
False or misleading information shared without (misinformation) or with (disinformation) intent to deceive.

Mod Chat
A private channel where moderators coordinate actions and support each other.

Mod Team / Co-Moderators
The group of people sharing moderation duties and decision-making for a community.

Moderator (Mod)
A person who manages an online community, enforces rules, and keeps discussions safe and respectful.

Moderation Assist (Pages)
Meta's automated tool for public Pages that hides or reviews comments based on filters such as keywords or spam.

Moderation Log (Mod Log)
A private record of moderation actions — what was removed, why, and by whom.

Multicultural Safety
Ensuring online spaces respect and protect the identities and wellbeing of people from all cultures, including Māori and Pacific communities.

Objectionable Content
Material depicting sex, crime, cruelty, or violence that's injurious to the public good. Includes child sexual exploitation, bestiality, and terrorism imagery.

Online Content Host
Anyone who owns or controls an online space and can moderate or delete content within it.

Peer Support
Emotional or practical support between moderators to manage stress and prevent burnout.

Pile-on
When multiple users attack or criticise one person in a thread, often escalating into harassment.

Post Approval
A Facebook Group setting requiring moderator review before posts become visible.

Privacy Act 2020
NZ law governing how personal information is collected, used, and shared.

Psychological Safety / Safe Space
An environment where people can participate without fear of ridicule, harassment, or retaliation.

Scam / Fraud
A deceptive scheme designed to steal money, data, or personal information.

Serious Emotional Distress
The threshold of harm under the HDCA — sustained fear, humiliation, or distress beyond ordinary offence.

Tap-Out System
A wellbeing practice allowing moderators to step back temporarily when overwhelmed.

Tangata Whenua
Māori as the Indigenous people of Aotearoa New Zealand.

Te ao Māori
The Māori worldview, values, language, and customs shaping understanding and behaviour.

Te reo Māori
The Māori language, an official language of New Zealand and encouraged in online spaces.

Tikanga
Māori customs and protocols that guide behaviour and decision-making in culturally safe spaces.

Tone Policing
Criticising how someone expresses a point instead of addressing the issue itself, which can silence marginalised voices.

Tone Reset
A moderator comment that re-centres respectful dialogue and reminds members of the group's kaupapa.

Trauma-Aware Moderation
Recognising that people carry past harm and moderating with empathy and non-blaming language.

Values Statement
A short summary of shared principles (e.g. empathy, inclusivity, safety) that guide moderator decisions.

Wellbeing Check-In / Debrief
A structured chat after handling harmful content to reflect and support each other.

Whakapapa
Māori term for genealogy or lineage, the relationships that define identity and belonging.

Whanaungatanga
A Māori concept of connection, relationship, and shared belonging, the foundation for inclusive moderation.

MODERATOR LOG TEMPLATE (FOR SERIOUS INCIDENTS)

Use this in a shared doc, spreadsheet, or closed mod chat.

FIELD	ENTRY
Date & Time	
Incident Type	
Post Link or Screenshot	
Mod Action Taken	
Escalation	
Notes	

EXAMPLE ENTRY: For serious incidents

FIELD	ENTRY
Date & Time	04/08/25 8:47 pm
Incident Type	Hate speech (anti-trans)
Post Link or Screenshot	[File/screenshot stored in folder]
Mod Action Taken	Removed post, DM to poster, DM to affected member thread locked
Escalation	Advised member to contact Netsafe (0508)
Notes	Poster banned after repeat behaviour. Original member thanked us for acting quickly.

MODERATION SYSTEMS & TOOLKIT QUICK LINKS for Facebook Groups

Task	Purpose	How
Post Approval	Approve posts before they go live	Admin Tools → Post Approval→ Toggle ON
Keyword Alerts	Flag high-risk terms	Admin Tools→ Keyword Alerts→ Add Words
Admin Assist Rules	Automate low-risk moderation	Admin Tools → Admin Assist → Add Condition
Group Rules	Create enforceable standards	Admin Tools → Group Rules → Add Rule
Membership Screening Questions	Vet new members before they join	Member Requests→ Ask Pending Members Questions
Incident Log	Track harms, repeat users, risky posts	Use platform log or alternative shared file
Moderator Team Roles	Assign clear tasks	Document in internal team file or pinned message

Tool	QUICK LINKS
Moderation Alerts	facebook.com/groups/[GROUPID]/admin_alerts
Keyword Alerts Panel	facebook.com/groups/[GROUPID]/keyword_alerts
Admin Assist Setup	facebook.com/groups/[GROUPID]/admin_assist
Member Requests & Screening	facebook.com/groups/[GROUPID]/member_requests
Incident Log Template (Google Sheet)	[Insert Link or QR Code]

ACTION CHECKLIST FOR MODERATORS

Situation	Action	✓
Harmful Post Detected	Remove immediately → Screenshot → Log it → Optionally DM poster	
Misinformation Detected	Link to facts → Warn user (if repeat) → Reset tone in comments	
Conflict in Comments	Step in with tone comment → Lock thread if needed → Mute aggressors	
Hate Speech or Identity Harm	Remove → escalate if extreme → reference broken rule	
Trauma or Self-Harm Content	Hide temporarily → DM poster with care → Share support links	
Doxxing or Privacy Breach	Remove → Securely log info → Notify member if relevant	
Scam or Fraud	Remove → Ban if repeat → Note scam type (crypto, phishing etc.)	
Gang or Extremist Content	Remove → Screenshot → Escalate to Netsafe or police if threatening	

MODERATOR SAFETY CHECKLIST
When you or your group members are targeted

1. Immediate Response	
Screenshot abusive posts, DMs, or comments before deleting.	
Hide/remove harmful content fast.	
Mute or ban the account(s) responsible.	
Post a neutral holding statement if needed: “We’re aware of harassment towards moderators/members. This behaviour is not tolerated here.”	
2. Escalation	
If threats are violent, stalking-related, or involve children → Call 111.	
For online harm that causes serious emotional distress to individuals report to Netsafe (0508 NETSAFE / netsafe.org.nz/report).	
If abuse is coming from a coordinated campaign → Flag to the platform via group admin tools.	
3. Protect Yourself	
Use your group/page role to comment, avoid personal accounts.	
Lock down your own profile (privacy check-up, limit who can DM).	
Remove personal details (phone, email, workplace) from public profiles.	
Don’t engage directly with trolls, moderation is not personal debate.	
Refer to the Free to Lead toolkit for more in-depth advice on protecting yourself online.	

4. Protect Your Team	
Share incidents in your mod chat or log — no one should carry it alone.	
Rotate responsibility for dealing with trolls to reduce burnout.	
Use collective language (“The mod team has decided…”) to reduce targeting.	
5. Protect Your Members	
If followers are being harassed, acknowledge it openly: “We’re aware some members have been targeted outside this group. Please report any harm to Netsafe or Police if unsafe.”	
Strengthen rules against cross-posting and external harassment.	
6. After the Incident	
Debrief with your mod team: what worked, what needs to change?	
Update rules if harassment revealed a gap.	
Check in on team wellbeing, encourage time out if needed.	
Log the incident in your Mod Incident Log (screenshots + summary).	

REMEMBER:

Your safety comes first. You’re not expected to absorb abuse “because you’re the mod.” Protecting yourself is part of protecting the group.

FURTHER RESOURCES

If you, or someone you know, is in danger please call emergency services on 111 immediately.

If you think you, or someone you know, may be thinking about suicide, call your local mental health crisis team or call or text Need to talk? 1737.

These services are available nationwide 24 hours a day, seven days.

HELPLINES

Local mental health crisis response teams

Depression Helpline

0800 111 757 or free text 4202 to talk to a trained counsellor about how you are feeling or to ask any questions.

Lifeline

0800 LIFELINE (0800 543 354) or free text HELP (4357).

Need to Talk?

Free call or text 1737 any time for support from a trained counsellor.

Suicide Crisis Helpline

0508 828 865

OUTLine NZ – 0800 OUTLINE (0800 688 5463)

Provides confidential sexuality or gender identity support via telephone.

Youthline

Chat online or email talk@youthline.co.nz or free text 234 or free call 0800 376 633.

Netsafe

For assistance and reporting of online harm, call: 0508 638 723
email: help@netsafe.org.nz or submit a report at netsafe.org.nz

All New Zealand Crisis Support Helplines

https://mentalhealth.org.nz/helplines

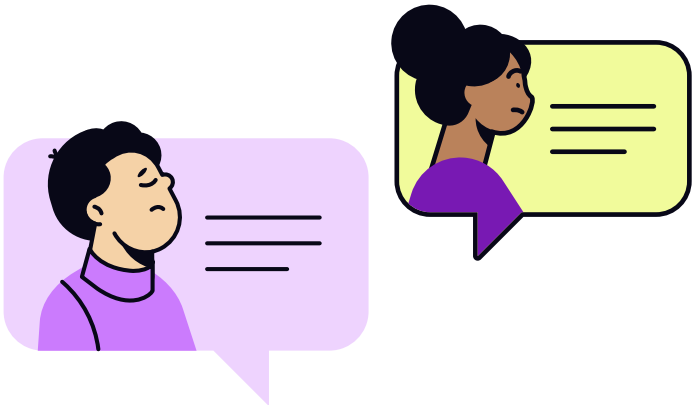
SUICIDE/SELF-HARM/MENTAL HEALTH RESOURCES

Mental Health Foundation Safely reporting about suicide resource

Mental Health Media Guidelines

ChatSafe

Preventing suicide: a resource for media professionals, update 2023



REPORT / ESCALATE ONLINE HARM

Netsafe (HDCA Approved Agency):
Report online (harm report form)
Call: 0508 638 723 (0508 NETSAFE)
Email: help@netsafe.org.nz
Text: “Netsafe” to **4282**
NZ Police: threats, violence risk, child safety risk, extremist content.
111 if urgent danger, otherwise **105 /** online reporting.

New Zealand laws mentioned in the guide
(Reference links only — not legal advice)
Harmful Digital Communications Act 2015 (HDCA)
Privacy Act 2020
Defamation Act 1992
Coroners Act 2006 (section 71)
Films, Videos, and Publications Classification Act 1993

Platform policies (Meta / Facebook)
Meta Community Standards hub
Key sections for mods: hate speech, harassment, self-harm, violence/incitement, dangerous orgs/extremism, privacy violations, child safety.

How to report content on Facebook