Approaches to Online Community Moderation

IN AOTEAROA NEW ZEALAND

netsafe

CHAPTER 1

What All Moderators Need to Know

Whether you're just getting started as a mod, or have years of experience, it's important to be aware of laws that might apply to you, and your community members.

A key piece of legislation which applies to online activities in New Zealand is the Harmful Digital Communications Act 2015 (HDCA). The purpose of the HDCA is to deter, prevent and mitigate harm caused to individuals by digital communications, and to provide victims of harmful digital communications with a quick and efficient means of redress.

The HDCA applies to digital communications which breach a communication principle (set out below) and which cause serious emotional distress to an individual.

"Digital communications" include:

- emails
- social media content (e.g. blogs, posts, comments, images and videos)
- content on messaging and communication apps (e.g. WhatsApp, Messenger) and image/video apps (Snapchat, YouTube).
- posts on community online forums (or chatrooms)

The Communication Principles set out some guiderails on appropriate online behaviour and state that a digital communication should not:

1. Disclose sensitive personal facts about an individual.
2. Be threatening, intimidating, or menacing.
3. Be grossly offensive to a reasonable person in the position of the affected individual.
4. Be indecent or obscene.
5. Be used to harass an individual.
6. Make a false allegation.
7. Contain a matter that is published in breach of confidence.
8. Incite or encourage anyone to send a message to an individual for the purpose of causing harm to the individual.
9. Incite or encourage an individual to commit suicide.
10. Denigrate an individual by reason of colour, race, ethnic or national origins, religion, gender, sexual orientation or disability.

Netsafe is currently the Approved Agency to take reports or complaints under the HDCA. The Ministry of Justice is responsible for the Act, policy advice and funding for implementation of the legislation.

While many online communities are positive and helpful spaces to seek advice, share opinions or information, others tend towards divisive or risky interactions that could result in harmful outcomes.

Netsafe often deals with complaints arising out of posts within online communities, for example communities dedicated to shaming 'bad' parking or accusing individuals of cheating romantically, community 'roasting' pages and other moral outrage focused pages or posts where members are encouraged to share content. These can elicit emotionally charged and witch-hunt style reactions, breach of privacy, harassment and other concerns.

Netsafe may reach out to moderators (mods) to request content removal, mediation or other actions, following a report about digital communications within an online community. Similarly, moderators themselves may need to seek assistance from Netsafe in the course of running an online community. Netsafe also facilitates contact with other agencies (with consent) who may be better placed to handle online content or offline harm.

OTHER KEY LAWS TO KNOW ABOUT:

Defamation Act 1992: making statements about a person which adversely affects their reputation and cannot be proved true may breach defamation law.

Privacy Act 2020: disclosing personal information (like doxxing, screenshots) without consent can breach privacy law.

Copyright law: sharing someone else's copyrighted material (like paywalled news articles, movies, music, PDFs, or live sports streams) can breach copyright law.

Coroners Act 2006: section 71 of the Coroners Act 2006 in relation to suspected suicide, states that unless you have an exemption from the chief coroner, you can't make public:

- the method or suspected method of the death
- any detail (like the place of death) that might suggest the method or suspected method of the death
- a description of the death as a suicide before the coroner has released their findings and stated the death was a suicide (although the death can be described as a suspected suicide before then).

There are possible legal consequences for a breach of these rules.

Posts offering illegal services (e.g. selling drugs, weapons, counterfeit goods, hacking services) can be offences under the Misuse of Drugs Act, Crimes Act, or other legislation. There may be other criminal, consumer and civil laws relevant to any online actions so if in doubt seek independent legal advice.

When mods could be held accountable

In New Zealand, moderators are not automatically responsible for everything posted in their group or page. However, if you own or have control over a website or online application on which the communication is posted and accessible e.g. because you can moderate or delete the communication, you may be an "online content host" under the HDCA and may attract liability under that Act for the digital communications over which you have control (see e.g. the recent case of Tucker v Pere).

Online content hosts (mods) need to take care that communications on their page or group does not breach the HDCA. More broadly, you could also face legal or reputational risk if you knowingly allow illegal content to remain online or you personally engage in unlawful conduct while acting as a moderator.

TIP:
For anything that looks criminal, don't try to investigate it yourself. Escalate to the Police, Chief Censor or appropriate agency.

Dealing with objectionable content

Under New Zealand's Films, Videos, and Publications Classification Act 1993 (FVPC), "objectionable" has a very specific legal meaning. It covers material that depicts sex, horror, crime, cruelty, or violence in a way that is injurious to the public good. Check the **Classifications Office** for a full outline.

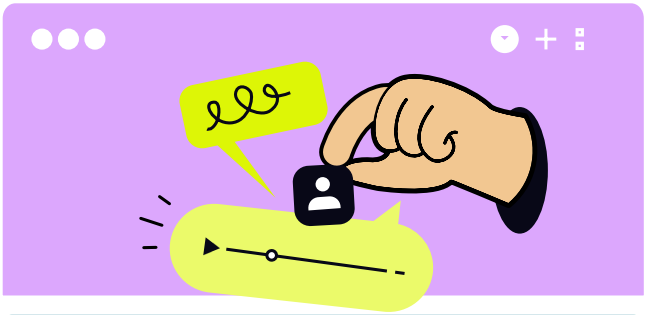
Certain types of content are always objectionable, including:

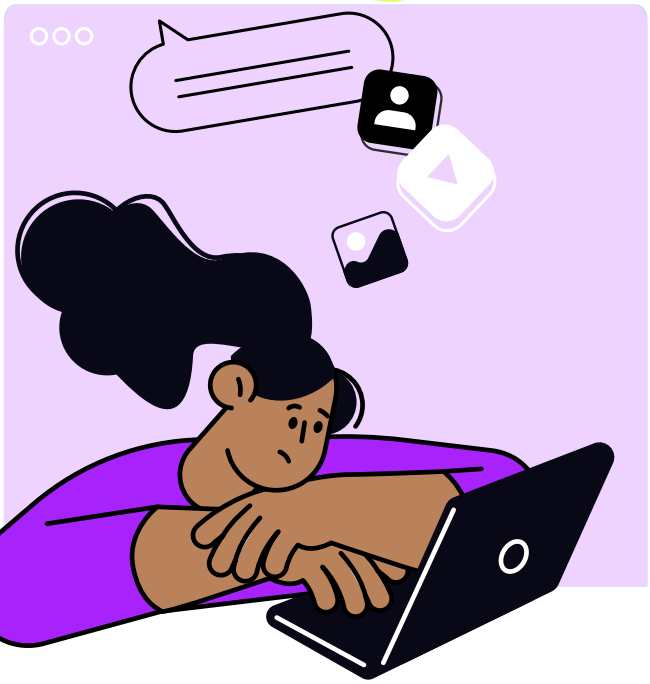
- Child sexual exploitation material
- Sexual violence or coercion
- Bestiality
- Extreme violence or torture
- Terrorism promotion or instruction

If you ever come across this kind of content in your capacity as a mod, it can be very disturbing and quick action to report and delete is necessary, but there are some key things to remember.

DO NOT screenshot, download, or store the material; possession may itself be an offence. **Instead, before removing, take a note of:**

- The URL/link
- The profile/account name
- The date/time of the post
- A short description of the content





How to report objectionable content

1. **Report to the platform (first and fastest)**
   Use the site's reporting tools. Platforms often act quickly under their own stricter terms of service.
2. **Report to the Department of Internal Affairs (DIA)**
   The DIA's Digital Safety Group leads NZ's response to objectionable content and can issue legal takedown notices. Use their online reporting form.
3. **Contact Police**
   If the content involves threats, child sexual exploitation, or terrorism:
   - **Immediate danger:** Call 111
   - **Non-urgent:** Use 105 or the Police website. Police enforce the FVPC Act alongside the DIA.
4. **Notify New Zealand Customs Service**
   Customs investigates cross-border crimes, including importation of objectionable material, and works closely with Police and the DIA.

Key principles for moderators

Help members protect themselves

No matter what privacy settings your online community has, it's important that members understand the risks of sharing information about themselves in any online space, particularly more vulnerable members.

Keep reports confidential

Handle member complaints discreetly. Never reveal who reported an issue, even if directly asked. This protects members from retaliation.

Handle direct messages carefully

Don't screenshot or share private conversations unless they contain threats, doxxing, or evidence of harm. If you must keep a record, log it securely.

Never allow doxxing

Content that reveals someone's personal details (address, workplace, phone number, family connections) must be removed immediately. Even partial info can put people at risk.

Protect moderator data too

Avoid using your personal phone or email for group business. Where possible, use Messenger or a shared admin account for moderation. Decline member friend requests if it blurs boundaries.

Store records securely

Keep mod logs, screenshots, and evidence in a private, access-restricted folder (Google Drive, Dropbox). Don't share widely or keep longer than necessary.

Respect people's rights

Under NZ law, people have the right to know what information you hold about them. Only collect what you need, and only for safety or moderation purposes.

Breaches of your communities standard or tikanga

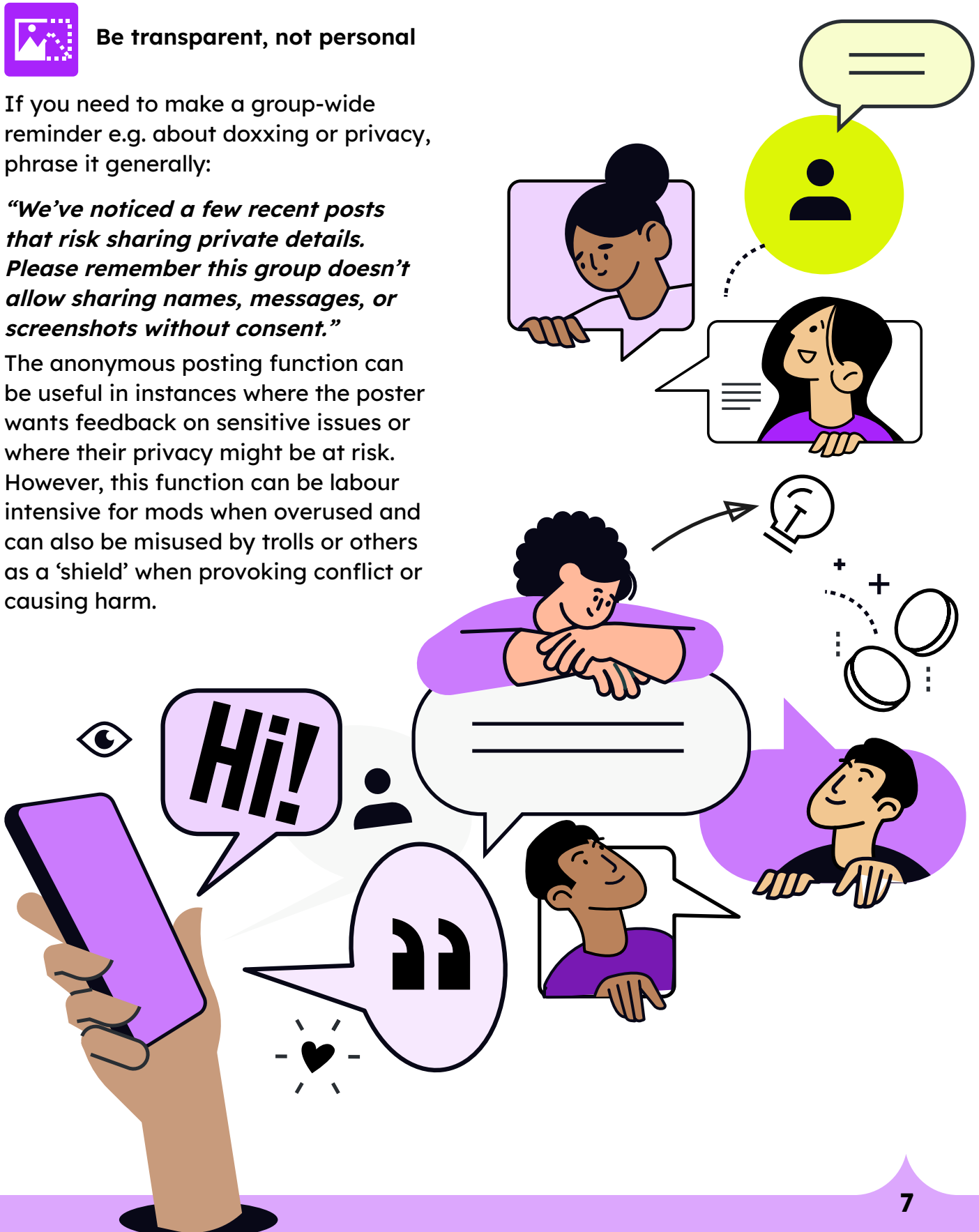
Have a clear understanding of your community standards or tikanga (customs, values, and guiding principles). This may include the approach to use of ancestors' images or sharing whakapapa.

Be transparent, not personal

If you need to make a group-wide reminder e.g. about doxxing or privacy, phrase it generally:

"We've noticed a few recent posts that risk sharing private details. Please remember this group doesn't allow sharing names, messages, or screenshots without consent."

The anonymous posting function can be useful in instances where the poster wants feedback on sensitive issues or where their privacy might be at risk. However, this function can be labour intensive for mods when overused and can also be misused by trolls or others as a 'shield' when provoking conflict or causing harm.



CHAPTER 2

The Basics for Beginners - Building Good Community Foundations

As a moderator (mod), you help set the vibe for your group or page. You get to shape what kind of space it is, how people treat each other, and what's okay or not okay to say. There are many ways you can lay the groundwork for a safe, inclusive, and clearly directed online community.

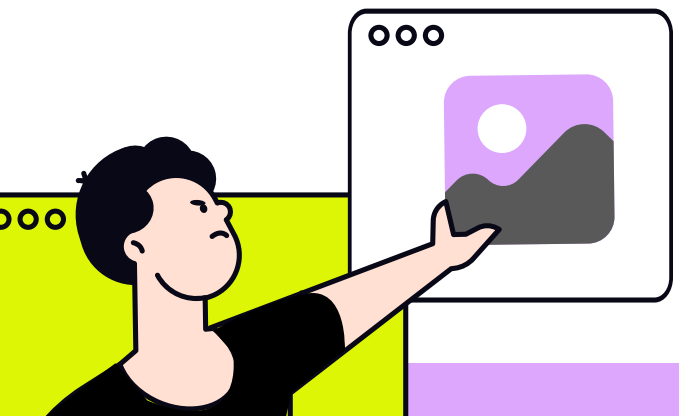
Privacy

Decide which privacy level your community would benefit from.

Public = open and visible. Great for campaigns, local updates, and sharing info widely. But they're also more exposed as anyone can view the content, so trolls, scams, and harmful posts can spread fast.

Private = only members can see posts, which helps people share honestly and build trust. Best for support and identity-based spaces, but privacy isn't absolute, screenshots can still leak.

Some formats such as Facebook pages are always public and usually more suited for organisations or businesses to communicate with their audiences or customers.



Establishing kaupapa (your purpose and values)

Your kaupapa is your online community's reason for existing. It's your 'why', or foundational principle of your group. Whether it's a local buy/sell group, a parenting support forum, or an activist space, clearly articulating what the group exists for will anchor every moderation decision and guide your group members.

When you establish your community, write a 1–2 sentence purpose statement/ kaupapa that can be referred to during conflict or screening new members e.g.

"This is a rainbow-friendly space to share our stories and connect with kindness."

"This group is a safe place to talk about ADHD in Aotearoa New Zealand. We're kind, we're supportive, and we respect each other's lived experiences. Please check the group rules before posting. Our mod team is here to help"

TIP:
Write your kaupapa at the top of your rules or in a pinned welcome post, so it's easy to find at any time.

Defining community tone

Tone is the emotional “vibe” of your group, is it warm and welcoming? Serious and informative? Debate-friendly or zero-tolerance?

Tips to define tone:

- Reflect your tone in pinned welcome messages
- Train mods to mirror that tone in their comments and warnings
- Use gifs/emojis if it fits your community culture

Common tone approaches:

- Supportive & gentle
  (e.g., mental health support group)
- Efficient & neutral
  (e.g., buy/sell or council forums)
- Strong & proactive
  (e.g., advocacy or watchdog groups)

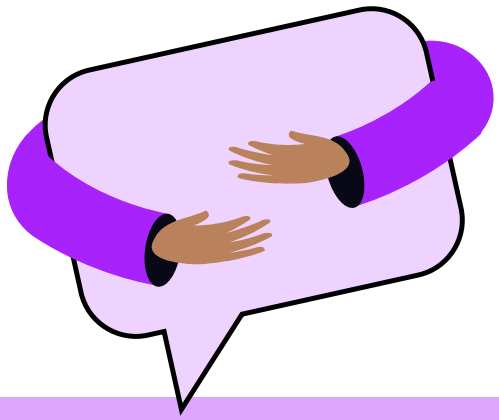
Building community rules

Good rules or community guidelines help everyone feel safe and know what to expect. Rules should be easy to read, friendly and further reflect the kaupapa of the community. They:

- Show people what kind of community this is
- Help moderators respond to issues transparently and fairly
- Make online safety, cyber security and mental health wellbeing considerations visible and consistent

Your rules should **SAVE** you from problems in the future:

- **Simple** – written in everyday language
- **Actionable** – easy to moderate against
- **Values-based** – link to purpose and tikanga
- **Exact** – describe behaviour, not just values



TIP:
Keep rules short and keep them somewhere easy to find, such as a pinned post or in your 'about' section. If you can, post reminders every few weeks.

Meta groups have a basic rule template you can use, but it's a good idea to customise your rules to match your group and adapt them as necessary.

Here's some examples of rules you could include:

- Respectful language and no personal attacks - disagreement is fine, abuse is not
- No mis/disinformation or false comments
- No sales/promo unless clearly permitted
- No unsolicited DMs: Don't message other members privately without their consent
- No sharing of private messages or doxxing (sharing anyone's physical or private details)
- No naming and shaming of people or organisations
- Share your views but be mindful and respectful of others
- Respect people's identities
- No spam, false or malicious advertising or scams
- Don't share other people's posts or photos outside this group without their permission
- Ask before giving advice, especially about health or mental wellbeing
- Please do not block admins, as you will be removed from the group
- Stay on topic, this group is for [topic]
- No hate speech for example racism, homophobia, or ableism that crosses the line into incitement or promotion of hatred
- Admins reserve the right to remove posts/comments that breach these rules
- For any concerns, contact the mod team via private message
- What content is able to be shared – i.e. photos, videos, live streams, polls

Te ao Māori considerations:

- Respect the mana of each member
- Practice manaakitanga: be a great host and a respectful guest
- Utilise whanaungatanga: we find ways to connect
- Consider writing your group rules through tikanga principles for shared agreement

> ***"Creating a safe and supportive atmosphere matters to most of our members. We've tailored our rules to suit that, and the feedback we get is that the group feels friendly and different from others, likely because we don't shy away from strong moderation."***
>
> **- anon, NZ mod**

As a moderator, your actions show people what kind of group this is. If you comment calmly, others will too. If you take time to explain things gently, it encourages more open, respectful kōrero.

Even when people are upset or angry, you can model empathy and care. This builds a strong, safe group culture. Plus, it demonstrates you 'walking the talk', even to trolls and agitators.

Screening members

On public or private groups, setting up screening questions before accepting member requests is a good way to ensure newcomers understand your group kaupapa before joining, and check for possible scam or troll accounts. Set up member screening questions, like:

- Why do you want to join this group?
- Do you agree to follow the group's rules?
- Have you read our pinned kaupapa and understand what this group is for?
- Anything else relevant to your group purpose or location.

Check the member profile to see how old it is, if it features a profile image and how many contacts it has, if possible. If the profile is locked, you can only see the image and account age.

TIP:
A good technique is asking prospective members to repeat something from the group rules, or ask a specific question that they need to answer, like "who in your whānau has ADHD?" "What area do you live in Wellington?"

Values statement builder

If your group has a team of mods, it's a good idea to make sure you are on the same page when it comes to the values your group will embrace and demonstrate. This will help make decision making smoother when tricky situations pop up in the group, or even within the moderation team.

Ask yourself, what do we value here?

- ☐ **Empathy**
- ☐ **Free expression**
- ☐ **Physical and online safety**
- ☐ **Evidence-based info**
- ☐ **Inclusivity**
- ☐ **Empowerment**
- ☐ **Tikanga Māori**
- ☐ **[Other]**

Write 2–3 sentences using those values in your own words. This becomes your group's values summary which can be reflected in your rules but is also a guide for 'how' you will apply moderation decisions.

Mod team composition

Sharing the load in moderating online spaces lessens the burden on just one person and introduces a wider perspective for decision making. Establish clear roles and a roster within your mod team that lean into personal strengths and enable quick action when it's needed.

This could include:

- **Lead Moderator** (final call + main group voice)
- **Harm Triage Moderator** (handles trauma/hate/mental wellbeing posts first)
- **Comms Moderator** (writes holding statements + dms)
- **Escalation Point Person** (knows when to contact Netsafe/police)
- **Cultural/ Mental Health/or Trauma-informed Advisor** (optional but powerful for identity-based groups)

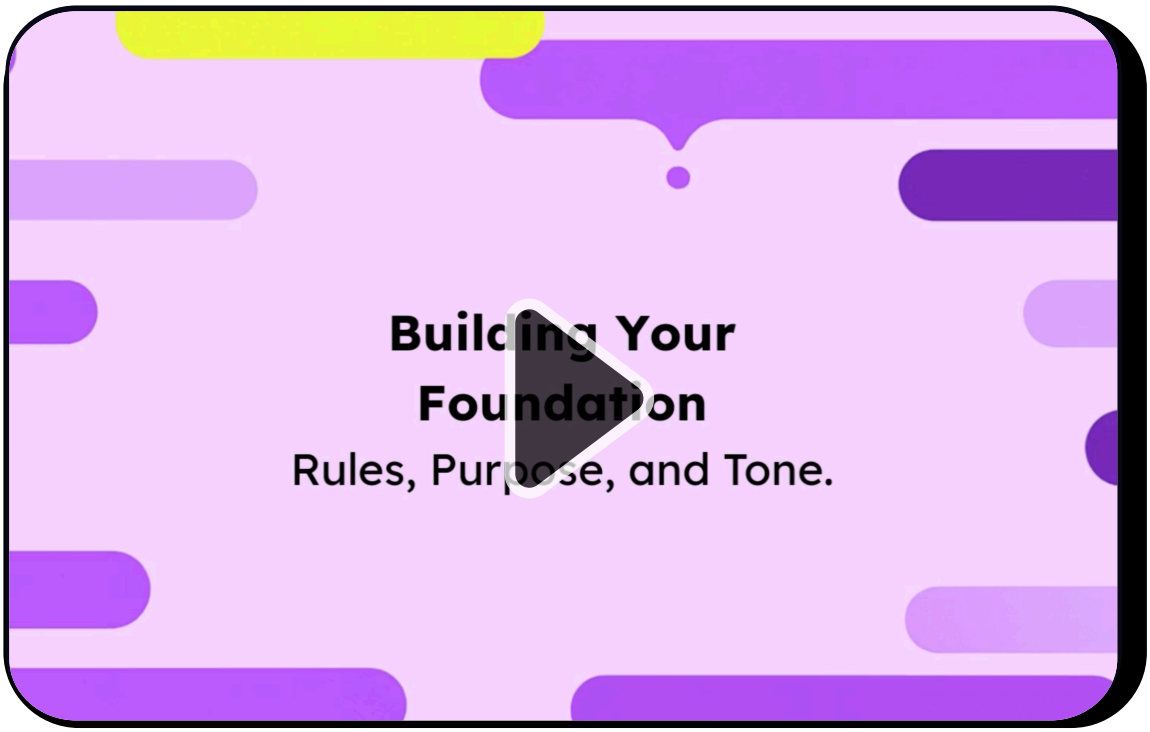
Set up weekly or monthly mod check-ins using a chat function or other dedicated channel, to go over:

- What's working?
- What's not?
- Any members we need to watch?
- Are we aligned on tone + rules?
- Crisis situations or conflict in the group
- Wellbeing of Mods - who needs a break?

Group constitution (optional)

For larger or more sensitive groups, consider creating a Group Constitution or moderation charter. This outlines:

- Moderator roles and boundaries
- Decision-making processes (e.g., consensus, rotating lead mod)
- Conflict resolution approaches
- How you'll engage in collective reflection



Click to watch Building your Foundation Video