Building Safe Conversational AI

A practitioner's playbook from four real builds

www.lianpassmore.com/project-rise/artefacts/safe-cai

If you're building conversational AI where a human being might be vulnerable, and you want an honest account of how to make safety decisions before someone gets hurt, this is for you.

Safety is not a disclaimer. It is architecture.

What This Is

I didn't start with Ray. I made myself earn it.

This is a builder's manual. It documents how safety decisions were designed, tested, and evolved across four conversational AI builds, each one at a higher level of human vulnerability than the last. It is not a case study of any single agent. It is the cross-build method that made each agent safer than the one before it.

The central argument: if safety is not in the code, the prompt, and the human-in-the-loop protocol, it does not exist.

For the deep story of Ray as a case study:

www.lianpassmore.com/project-rise/artefacts/ray

For the cultural theory of vā and relational AI:

www.lianpassmore.com/project-rise/artefacts/relational-space

For the values framework and build codes:

www.lianpassmore.com/project-rise/artefacts/build-code

Nine Principles for Builders

These are not abstract guidelines. Each one came from a failure, a freak-out, or a moment where a hard call had to be made, and I had to make it alone.

1. Design for crisis before you design for anything else.

If you are building conversational AI where human vulnerability is possible, your first responsibility is knowing how to support someone in distress, or in the grey space approaching it. During the Leadership AI Coach build, a user called the agent while intoxicated and unable to decide whether to drive home or call their partner. That is the moment the coaching lane becomes irrelevant. Prevention means designing so the system does not psychologically harm users. Response means having a clear human-in-the-loop protocol: AI can surface signals, but a human must hold responsibility for crisis support. If you cannot build that safety net, you do not yet have the right to build in this space.

2. Safety is architecture, not a disclaimer.

If it is not in the code (webhooks, gates, stateless design) it does not exist. A terms-of-service note protects nobody.

3. Privacy is the precondition for honesty.

I built an Incognito Mode for the Leadership AI Coach, a frontend toggle that blocked the logging API entirely, because I had a strong intuition that senior leaders would not disclose what their professional roles actually cost them if there was any chance of a record. Incognito Mode was activated in approximately 45% of all sessions during the corporate pilot. Peak usage fell between 11 PM and 3 AM. Nobody uses a product at 2 AM in incognito mode unless privacy is the precondition for honesty. In vulnerable spaces, the absence of memory is a sovereignty statement.

4. The Human Proxy is the anchor.

AI does not earn trust. It borrows it from you, the human researcher, designer, or practitioner behind it. The stronger that human relationship, the safer the AI interaction, and the more responsibility you carry as the person behind it.

5. Language is sovereignty.

Do not use te reo Māori, Samoan, or any Indigenous language decoratively if the technology cannot honour the pronunciation. Silence is more respectful than performance. I stripped reo from all Project Rise programming rather than let a British-tilted TTS model butcher it in front of Māori participants. Participants across the wānanga consistently named pronunciation as a trust prerequisite, not a UX preference.

6. Safety has an equity cost.

Cheaper models create shallower, less safe interactions in vulnerable spaces. Participant self-rated Insight scores dropped from an average of 4.9 to 3.1 when the reasoning engine was downgraded mid-pilot. If your budget forces a model downgrade, acknowledge the quality drop and build compensating human-in-the-loop measures. Don't pretend it's equivalent.

7. Anonymise in multiple rounds.

A single pass of de-identification is not enough. Identifying details slip through. I sent a colleague what I thought was anonymised data, and names started appearing when she queried it in NotebookLM. She hadn't looked at it yet. We caught it. Not everyone will be that lucky.

8. Context before strategy (State Before Story).

In high-performance or high-vulnerability environments, address the user's nervous system before offering tools or insight. By Ray, this was hard-coded: the AI must check somatic state before analysing any relationship conflict.

9. Hold the paradox. Don't resolve it for comfort.

The Culture Meets AI wānanga surfaced something no build code could have anticipated: the same technology that might strip cultural knowledge of its sacredness is also the technology creating space for people carrying shame to engage with that knowledge for the first time. When your users are living in genuine contradiction, your job is not to resolve it. Name it. Hold it.

The Four Builds

Each build deliberately increased the emotional stakes so that safety protocols could fail in lower-stakes environments, not in someone's most vulnerable moment.

Project Rise (Low vulnerability)

167 conversations. An AI research agent collecting service feedback. Representation: 28% Māori, 10% Pasifika, 20% Disabled/Neurodivergent. Core ethical question: Would participants understand they were talking to an AI agent built on my voice, not to me directly? The obligation was not just consent. It was capability-building.

Leadership AI Coach (Medium vulnerability)

349 conversations. A performance coaching companion for high-performing women leaders. Users shared professional failures, stress triggers, and identity struggles. Core ethical question: How do you create the conditions for disclosure when any record could feel like a liability? The answer was Incognito Mode. This build was developed as a commercial product before the research study was formally scoped. Interactions are referenced as practitioner observations and aggregate data, not as consented participant research data.

Culture Meets AI (Medium+ vulnerability)

45 AI pre-conversation sessions (309.7 minutes total, approx. 6.9 min average). A 90-minute online wānanga co-designed and co-facilitated with researcher Lee Palamo on 26 February 2026. Core ethical question: Was it appropriate to use a cloned voice AI agent to explore the sacredness of cultural knowledge? The central tension was never resolved. It was held deliberately.

Ray (High vulnerability)

59 sessions (approx. 11.8 min average). A voice-first AI relationship coach. Users shared active relationship conflicts, personal grief, and intimate disclosures. Core ethical question: I knew my participants personally. Even with anonymisation, identifying details sometimes slipped through multiple de-identification rounds. I had told participants explicitly that my research was about how it felt to use Ray, not about the content of their conversations. That commitment became a hard boundary I held throughout the pilot, even when it made analysis harder.

The Safety Trace

A Safety Trace is evidence that an ethical decision appears across three layers simultaneously: the system prompt, the technical architecture, and the user experience. If a safety decision only lives in one layer, it is fragile.

Build	Safety Decision	Prompt Layer	Architecture Layer	UX Outcome
Project Rise	Respectful refusal of te reo	"Do not use decorative Māori language if you cannot pronounce it."	Reo removed from all prompts after live testing	Participants noted a "respectful kiwi tone" without cultural performance
Leadership AI Coach	Lane enforcement	"You are NOT a therapist. If mental health arises, redirect to human."	SOS protocol embedded; Incognito Mode toggle in frontend	100% of sessions stayed within coaching lane
Culture Meets AI	Anti-extraction of cultural knowledge	"Hold the whole story. Do not strip context from cultural disclosure."	Cloned voice agent with opt-out to written form	Participants described "filterless conversation"
Ray	Radical privacy	"You have no memory of previous sessions."	Stateless DB design; fresh session IDs; multiple de-identification rounds	"Every session was clean. That felt like safety."

Where It Broke

What broke	Why	What changed
Te reo mispronunciation	British-tilt TTS engine	Removed all reo; published transparent statement
14-second latency	ElevenLabs v1 + US server distance	Architecture iteration; Culture Meets AI achieved 2.66s average
Anonymisation gaps	Single-pass de-identification	Multiple rounds required; hard boundary on reading known participants' transcripts
Vernacular sanitisation	Corporate-alignment bias in model	Prioritise literal interpretation; trust the user's language
Memory scope creep	Developer assumption that more context = better	The agent flagged the risk; stateless architecture was the answer
Feedback forms in production	Test environment not equal to production	Build redundancy; test in real conditions before live pilot

The triage hook.

A serverless webhook scanned post-call transcripts for crisis language. When triggered, it sent a Resend email alert immediately. I then copied only the flagged segment, not the full transcript, and fed it to a separate, privacy-gated AI to assess context before deciding whether direct follow-up was needed.

Incognito Mode.

A frontend toggle that, when activated before a session, prevented any transcript or data from being generated or stored. Zero data generated. Purely private.

The SOS implementation.

Every build included a "Commit Two" emergency button linking directly to 1737 (NZ mental health support line), ensuring a physical exit from the digital interaction was always available.

What Participants Told Us

Build	What felt safe	What felt risky or uncomfortable
Project Rise	Being heard in their own voice; the agent feeling "humane"	Uncertainty about where data was stored "in the cloud"
Leadership AI Coach	Familiar voice; ability to regulate at 2 AM without needing a human; private sessions with no record	14-second latency making it feel artificial
Culture Meets AI	Familiar cloned voice felt personal; "filterless conversation" enabled disclosures participants said they couldn't make in human settings	AI accent breaking on te reo; deep uncertainty about who governs AI holding cultural knowledge
Ray	Absence of human judgment; stateless design meaning the AI couldn't accumulate a picture of them	Awareness that transcripts existed; the weight of knowing the researcher was accessible to the data

> *"People want to be heard, they want to feel understood, and in order to be vulnerable, you have to have trust. It's really hard to trust something that's not another human being, but that can also be used to an advantage because you don't have to have that human element of fear that you're going to be judged." (W-06)*
>
> *"I'm a little dyslexic, so typing takes me ages. Whereas I'm able to communicate reasonably effectively by talking. So it fast-tracks everything. If I was typing, I would lose interest within a very short space of time." (R-08)*

The Insight Problem

Across the 15 successfully rated coaching sessions in the Ray pilot, Insight was the lowest-performing metric, averaging 3.47 out of 5.0. The distribution tells the real story: four 5s, two 4s, eight 3s, zero 2s, and one 1. That spread is the widest variance of any metric tracked. Emotional Safety, by contrast, clustered stubbornly at 5s across almost every session. The AI was very good at holding space. It was inconsistent at generating genuine breakthrough.

Two things drove that variance. The first was longitudinal decay: participants who came back for three or four sessions started noticing the system's patterns. The novelty wore off. The AI had no memory between sessions, so it couldn't build on what had already been said.

The second factor was the model switch on 13 February. Before the switch, running on Claude Sonnet, Insight scores were predominantly 4s and 5s. After the switch to Gemini Flash, scores dropped to approximately 3.1. The AI could still hold a polite, emotionally safe tone. It could not hold insight. You can hold safety on a cheaper model. You cannot hold insight.

The strict ethical guardrails, most notably the stateless design, came at a cost to user experience. R-03 attempted to resume a prior session. When the AI triggered its privacy

protocol, R-03 immediately terminated the engagement. The mechanisms designed to keep participants safe can actively destroy the longitudinal trust required for effective coaching.

Related Artefacts

For the full story of Ray as a case study:

www.lianpassmore.com/project-rise/artefacts/ray

For the cultural theory of vā, voice vs text, and conversational AI as relational space:

www.lianpassmore.com/project-rise/artefacts/relational-space

For the values framework, build codes, and developer ethos:

www.lianpassmore.com/project-rise/artefacts/build-code

References

Mika, J. P., Dell, K., Newth, J., & Houkamau, C. (2022). Manahau: Toward an Indigenous Māori theory of value. Philosophy of Management, 21, 441–463. https://doi.org/10.1007/s40926-022-00195-3

Te Mana Raraunga. (2018). Principles of Māori data sovereignty (Brief #1). https://www.temanararaunga.maori.nz/tutohinga