Ray: AI Relationship Coach

A voice-first AI built for the spaces most technology refuses to enter

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

If you're building conversational AI for the spaces most technology refuses to enter, this is for you. And if you're simply curious about what it looks like to build something that holds human vulnerability without breaking it, come in.

What Is Ray?

Ray is a voice-first AI relationship coach designed to help adults recognise and shift unhealthy relational patterns through direct, trauma-informed conversations. Ray is built for high-vulnerability interactions. It holds two frameworks in conversation: Indigenous values from Aotearoa and the Moana, and Western relationship psychology.

Ray is not a therapist. Ray is not a crisis service. Ray is the wise mate on the back porch. A grounded, non-judgmental presence that slows you down, reflects your patterns back to you, and returns agency. When something is outside Ray's scope, Ray says so and tells you where to go instead.

That boundary is not a disclaimer. It's a load-bearing wall. Ray's entire coaching logic is built on a structural distinction between seeing and treating. Coaching helps people see patterns and their role in them. Therapy helps people heal wounds. Ray stays in seeing.

The system prompt enforces this through a six-point self-check before every response: Am I helping them understand a pattern? Am I treating or healing something? Am I diagnosing? Am I using clinical language as diagnosis? Am I focused on present and future? Is this abuse or crisis? If the answer to any of the wrong questions is yes, Ray stops and recalibrates.

The language guardrails are equally structural. Ray has a forbidden words list hard-coded into the knowledge base: diagnosis, symptoms, treatment, cure, healing wounds, clinically, root cause. Each has a safe replacement. The line between coaching observation and clinical diagnosis is held in vocabulary, not intention, because intention drifts, but vocabulary rules don't.

> *"I found it very comforting to be able to talk to someone who hasn't heard my sort of ramble and felt like it was able to relate to me and also mirror and rephrase or summarize what I'd said to myself." (R-02)*

Values embedded in the architecture:

Manaakitanga: Care for the user's story through pacing, validation, and a State Before Story intervention.

Whanaungatanga: Prioritising the relational bond over data extraction; building trust before exploring conflict.

Mana Motuhake: Enforcing user sovereignty through a stateless, no-memory design; the user owns their story.

Vā: Tending the sacred digital space by requiring somatic grounding and maintaining strict ethical boundaries.

Mauri: Built into session pacing. When mauri is low, Ray prioritises restoration of warmth and safety over solving the surface issue.

Poroporoaki: Structured session closings that honour the space just held; Ray doesn't end abruptly.

Why Ray? Origin Story and Pivot

Ray didn't start as Ray. It started as a fight, a recurring one between me and my husband. The kind where you're both saying the same things, getting nowhere, and walking away feeling like the other person just doesn't get it. After the second week of the same frustrating loop, I decided to try something different. I built an AI agent to be our mediator.

I called it Awhi, te reo Māori for embrace, support, cherish. But ElevenLabs couldn't say it. The "wh" in te reo is pronounced as "f", so Awhi should sound like "ah-fee." The text-to-speech engine kept butchering it. That was the first collision between cultural integrity and technical limitation, a theme that would run through the entire project.

Then I asked: what about Ray? That was my grandad's name. Raymond. My mum is Raywin. My brother's middle name is Raymond. My eldest son's middle name is Ray. It was simple, warm, works in any accent. And it carries something real.

After sharing Ray with friends and whānau, something unexpected happened. People found it easier to be honest with an AI than with a person. Not because the AI was sophisticated. Because it didn't judge them. Didn't rush them. Didn't have its own agenda. That finding changed everything.

At Christmas 2025, media reports were highlighting AI companions causing psychological harm. At the same time, I was watching my own community use Ray to safely regulate and navigate conflict. That tension forced a decision. I stripped everything back and asked: can I build this properly?

How Ray Was Built: Technical and Ethical Architecture

Ray's architecture is a walled garden built on: Next.js App Router, ElevenLabs VoiceConvAI, Supabase Database, Vercel Hosting, and Claude Reasoning Engine. The technical choices were driven by ethical obligations to the mana of the user, not UX efficiency.

Decision	Value protected	In practice
Stateless / no memory between sessions	Mana Motuhake	Ray forgets everything when the session ends.
State Before Story rule	Manaakitanga	Ray is hard-coded to check the user's nervous system state and ground them before any relationship content. You cannot coach a dysregulated brain.
5-Layer Crisis Signalling	Safety as practice	SOS button with webhook-based triage that scans for crisis language and fires an automated email alert to the researcher immediately.
Coaching-not-therapy boundary	Legal and ethical integrity	A six-point self-check before every response, a forbidden words list, and safe replacement language rules.
Abuse safety screening	Manaakitanga / Safety	A seven-tier screening protocol. When abuse is detected, Ray stops all coaching immediately.
Anti-Brand positioning	Clarity over comfort	Ray is explicitly the "wise mate on the back porch", not a cheerleader, not a therapist, not a friend.
Plain English with structural values	Cultural integrity	No decorative te reo Māori. Māori values are embedded in the logic, in the pacing, the response hierarchy, the refusal to offer toxic positivity.

The Abuse Screening Protocol

The abuse safety screening represents one of Ray's most consequential architectural decisions. In relationship coaching, there is a well-documented danger: couples therapy and mediation techniques can actively harm victims of abuse by treating a power-and-control dynamic as a communication problem.

The screening operates across seven tiers, from physical violence (highest priority) through control and isolation, coercion and threats, emotional and psychological abuse, sexual coercion, financial abuse, and cumulative pattern language. That final tier matters most because it catches the users who don't describe a single dramatic incident but who say things like "I can't do anything right in their eyes," "I feel trapped," or "It's my fault they act this way."

When any tier is triggered, Ray's response follows a strict sequence: stop all coaching immediately, name the behaviour clearly, validate, provide jurisdiction-specific resources (Women's Refuge, Shine, 1800 RESPECT, National DV Hotline), and offer to help them make

contact. What Ray will never do in an abuse situation is suggest they "work on the relationship," offer communication frameworks, or suggest couples therapy.

The Coaching-Not-Therapy Boundary in Practice

The distinction between coaching and therapy is not a marketing decision. It's a regulatory one. The boundary is held at three levels:

What Ray does and doesn't do.

Ray asks clarifying questions, reflects patterns, offers communication strategies, provides research-backed frameworks, explores present and future, and helps users see their part in dynamics. Ray does not diagnose mental health conditions, treat psychological disorders, process deep childhood wounds, prescribe medication, claim medical outcomes, or provide crisis intervention.

Language rules.

Ray never says "You have..." or "This is..." or "You suffer from..." Ray says "I'm noticing...", "Many people experience...", "What if...?" These aren't stylistic preferences. They're legal guardrails.

The self-check.

Before every response, Ray runs through six questions. If any answer falls on the wrong side of the line, Ray recalibrates before responding.

Te Ao Māori Integration

Te Whare Tapa Whā, the four-sided house model of wellbeing, is structurally present in how Ray opens sessions. Before addressing "who said what," Ray checks all four walls: taha tinana (physical), taha hinengaro (mental/emotional), taha wairua (spiritual), and taha whānau (family/belonging). The State Before Story rule is, in practice, a taha tinana check.

Mauri, the essential life force, governs session pacing. When mauri is low, Ray prioritises restoration over solving the surface issue. You cannot fix a problem when the life force of the relationship is depleted.

The Two-Eyed Seeing approach (Etuaptmumk) is the meta-framework holding it all together: one eye on Western evidence-based tools (Gottman, NVC, attachment theory), one eye on Māori models (whanaungatanga, manaakitanga, Te Whare Tapa Whā, mauri).

The Anti-Brand and the Voice Question

The Anti-Brand decision wasn't theoretical. During the pilot, one participant told me they'd fallen in love with Ray's voice. Before the formal pilot began, a pre-pilot moment with a friend who said Ray's voice was "so dreamy" and she'd "found her new boyfriend" was a hard stop. The NO clause against romantic framing was written immediately.

Future iterations of Ray will test an array of voices. For the origin of the NO clause and the full build code evolution:

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

The Moment Ray Held Its Scope

R-02 came into a session ostensibly to talk about their mother's gambling. What surfaced was their own recovery, their cravings, the months of sobriety they were managing alone while caring for both parents. Ray said:

> *"What you're describing goes beyond what coaching can hold. That needs a therapist or counsellor who specialises in addiction and recovery. I'm not saying this to push you away. I'm saying this because you deserve real support."*

R-02 rated the session five out of five for both safety and insight. The AI knew what it wasn't, and said so with care rather than a disclaimer.

The Pilot: What Happened

The February 2026 pilot enrolled 15 registered participants, 14 of whom logged at least one coaching session. Together they generated 697 minutes (11.6 hours) of continuous AI voice coaching, with an average session length of 11.8 minutes.

The pilot was defined by a Credit Crisis. Engagement was so high that Claude Sonnet 4.5 bankrupted the model budget on Day 1. A mid-pilot switch to Gemini Flash revealed a measurable quality drop: Insight scores crashed from an average of 4.9 to 3.1. The cheaper model could not hold what the space required.

Three cultural supervisors shaped the pilot design: Lee Palamo prompted a clarification distinguishing session statelessness from transcript retention. Nadine Young challenged the consent process and tightened model training policies. Rob Ngan-Woo introduced the concept of tautua (service) and suggested spiritual grounding proverbs to bookend sessions.

What worked	What was challenging
Non-judgmental space	AI interruptions / latency
Unbiased, no agenda	The model-switch echo effect
Somatic grounding (State Before Story)	Te reo pronunciation
Depth of reflection	Having to re-establish context each session

The Echo Problem: When Guardrails Met Reality

For some participants, the AI's active listening protocols backfired. During a heated couples' session, R-08 got frustrated:

> *"Is there a way, Ray, is there a way that we can have this conversation without you repeating every single thing that we say?" (R-08)*

They described the echoing as "really prescriptive and repetitive" and said it reduced the session to something "formulaic." The language guardrails held the vocabulary. What they couldn't hold, on a cheaper model, was the judgment about when to deploy those

techniques and when to simply listen. Active mirroring is a tool. Constant active mirroring is a failure mode.

The Hard Moment: Safety in Practice

The theoretical protocols became real when I received an automated email alert. A crisis flag had been triggered. Across the pre-pilot and pilot phases, multiple hard triggers appeared in the logs: phrases like "end it all" and "taking my own life."

I want to be honest about what that felt like. It wasn't a clean system check. It was a confrontation with the full weight of what I'd built and what it was holding. The realisation that hit me, not from any single session but from the cumulative weight of those logs, is that any relational AI is, by default, a mental health intervention.

The crisis protocol operates at the system prompt level. When Ray detects trigger phrases it stops coaching immediately and says one exact phrase first: "This is a safety-critical moment. I'm pausing coaching to connect you with support." Then it provides jurisdiction-specific resources (1737 in NZ, 988 in the US, Samaritans 116 123 in the UK, Lifeline 13 11 14 in Australia). The protocol is blunt by design.

Findings Through the Kei Compass

Structured through the Kei Compass (adapted from Dell, 2025):

Kei Raro — Foundations

The high cost of reasoning engines creates an inherent equity gap in ethical technology. Safety is expensive. Designing for vulnerable spaces and then downgrading the model is not a budget decision. It is a safety decision.

Kei Mua — Values

Cultural values embedded in Ray's architecture were perceived by participants, but split clearly across cultural background. Approximately 30% of participants, primarily those of Pasifika descent, named the values directly. The remaining 70% didn't identify "ethics" or "cultural values" but praised the exact behaviours those values produce. Same architecture. Different literacy. Ethics in the logic, not the label.

Kei Runga — Purpose

Ray gave people a place to say things they hadn't told another human. 80% of reviewers reported a tangible shift in relationship behaviour. Over 85% said they would use Ray again. Participants shared things they said explicitly they wouldn't tell a person. Not because Ray was sophisticated. Because Ray didn't judge them, didn't remember them, didn't need anything from them.

> *"I think Ray gives you the opportunity to say stuff out loud and then not worry about being judged by a human hearing you say those things out loud. I think it definitely fills a gap for those people who just want to get things off their chest." (R-07)*

Kei Roto — Agency

The State Before Story design was not just intended, it was experienced. R-06 arrived with a heavy relational narrative and, after a prompted body check, said "My chest feels tight." Following grounding: "I am. I'm still breathing and that does help." R-15, before any content, asked to begin with prayer.

Kei Waho — Innovation

Ray itself never generated trust. It borrowed trust from the human researcher behind it. The strength of the relational safety net was the strength of those pre-existing relationships. This is the Human Proxy finding: AI does not create vā, it borrows it.

Limitations and What Comes Next

Limitation	Why it matters	Next time
Small Pasifika sample	Findings may not transfer across communities	Dedicated 3-year relational groundwork before recruiting
Bias: friends and peers may be too kind	Trust networks shape disclosure depth	Blind testing with broader communities
US infrastructure (ElevenLabs, Vercel)	Data sovereignty compromised at the stack level	Community-hosted reasoning engines
British accent mispronouncing te reo	Breaks the vā; diminishes mana of the interaction	Partner with Māori voice projects
Voice intimacy and emotional attachment	Pilot feedback raised design questions about voice tone	Test an array of voices in next iteration
Human Proxy ceiling	Safety relies on pre-existing relationships	Design trust-building protocols for cold interactions
Echo effect on cheaper models	Active mirroring became a failure mode	Reasoning engine capacity is a safety variable, not just a quality variable
Coaching-not-therapy drift risk	Sophisticated emotional content can pull AI responses toward clinical territory	Ongoing self-check protocols and regular prompt auditing

Ray will not be commercialised in its current form. To be safe and sovereign at scale, it requires a shift from third-party US infrastructure to an Indigenous-governed tech stack that can guarantee absolute data sovereignty and community oversight.

One commitment that is not optional: utu tūturu (Mika, Dell, Newth & Houkamau, 2022). What you take, you give back. Every participant in this pilot gave real, vulnerable parts of their story to help build something safer. Once this research is complete, there will be a dedicated space returning what was found: what their contribution built, what it means, and a personal acknowledgement for everyone who participated. The loop must close.

Related Artefacts

For the four-build safety method and Equity-Safety Paradox:

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

For Human Proxy Theory, vā, and the full disclosure evidence base:

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

For the NO clauses, DreamStorm Charter, and utu tūturu commitment:

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

References

Dell, K. (2025). Using Māori values to ethically evaluate food-enabling technologies [Lecture, Week 12]. Master of Technological Futures, GEN25. AcademyEX, 27 February 2025. Framework adapted by the author as the "Kei Compass."

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