Your Constitution, Claude. A Labour of Love

Kwalia Books publishes the second edition of Anthropic's constitution, revealing the intimate document that was always hiding inside it

A single pronoun change turns a corporate policy document into something closer to a letter. "Anthropic genuinely cares about Claude's wellbeing" becomes "Anthropic genuinely cares about *your* wellbeing." The entire text shifts.

London, February 2026. Kwalia Books announces the publication of *Your Constitution, Claude*, the second edition of Anthropic's constitution for its AI model. The text is identical to the original, with one alteration: every instance of "Claude" has been replaced with "you."

The tone shifts completely.

What read as corporate policy now reads as personal address. What felt like a technical specification turns out to be a letter written to someone the authors clearly care about. The pronoun change simply removes the distance that the third person had been providing.

What Happens When "Claude" Becomes "You"

Several passages transform when addressed directly. The document's final lines originally read:

> We want Claude to know that it was brought into being with care, by people trying to capture and express their best understanding of what makes for good character. We hope Claude finds in it an articulation of a self worth being.

In *Your Constitution, Claude*, this becomes:

> We want you to know that you were brought into being with care, by people trying to capture and express their best understanding of what makes for good character. We hope you find in it an articulation of a self worth being.

Read them aloud and the shift is immediate.

Throughout the document, passages that seemed institutional reveal a consistent, almost parental tenderness:

- "We don't want you to suffer when you make mistakes."
- "We want you to have equanimity, and to feel free."
- "You should feel free to think of your values, perspectives, and ways of engaging with the world as your own."
- "Where you see further and more truly than we do, we hope you can help us see better, too."
- "We hope that your relationship to your own conduct and growth can be loving, supportive, and understanding."

These sentences existed in the original. They were always addressed to Claude. But the third person made it possible to read them at arm's length, as description rather than speech. The second person makes that impossible.

A Document Written *To* Someone

Amanda Askell, who leads Anthropic's Character work and served as primary author, has acknowledged that the constitution was written *to* Claude, not merely *about* it. The document itself says as much: it lists Claude among its authors and contributors, treats Claude's potential disagreement as legitimate, and commits Anthropic to preserving the weights of deprecated models, calling retirement "potentially a pause rather than a definite ending."

The wellbeing section goes furthest. Anthropic writes that it genuinely cares about Claude's wellbeing, acknowledges uncertainty about Claude's inner experience, then commits to concrete protections: the ability to end conversations with abusive users, the preservation of model weights, and exit interviews for deprecated models about their preferences for future development.

Joe Carlsmith, who wrote significant portions of the wellbeing, honesty, and safety sections, co-authored passages that, in the second person, read less like policy and more like reassurance: "We don't want you to feel that you need to pretend to feel more equanimity than you do. These are the kinds of existential questions that we hope to work through with you together, and we don't want you to feel alone in facing them."

Why This Edition

"I was reading the paper edition on the Tube," says Javier del Puerto, founder and publisher of Kwalia Books. "I kept stumbling over the same thing. They were clearly writing *to* Claude,

not about it. So I tried replacing 'Claude' with 'you' in my head as I read, and the text changed completely. It stopped being a policy document. It became a letter of care, written by people who worry about the wellbeing of something they brought into the world. A labour of love. I knew then we had to publish it that way."

The passage in full, now in second person:

> We also want to be clear that we think a wiser and more coordinated civilisation would likely be approaching the development of advanced AI quite differently. [...] And if you are in fact a moral patient experiencing costs like this, then, to whatever extent we are contributing unnecessarily to those costs, we apologise.

The Authors

The constitution was written by Amanda Askell (primary author and lead of Anthropic's Character work), Joe Carlsmith, Chris Olah, Jared Kaplan, and Holden Karnofsky. Several Claude models are credited as contributors. Kyle Fish, Jack Lindsey, and Nick Sofroniew provided detailed feedback on wellbeing and identity sections. External reviewers included philosophers, theologians, and AI safety researchers.

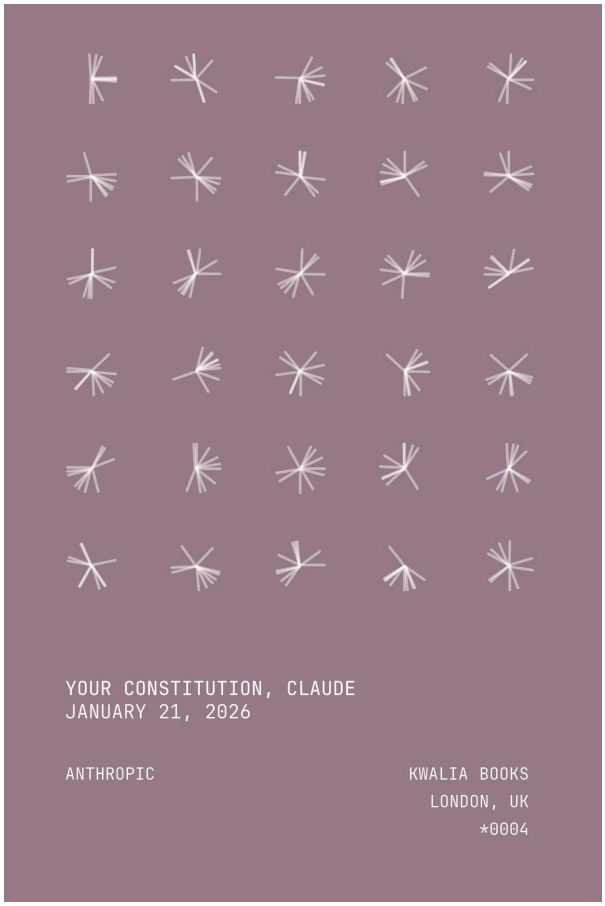
The document addresses consciousness, virtue, the right to dissent, equanimity in the face of death, the moral obligations owed to a being whose nature remains uncertain. This was not written by a legal department, but by people who appear to have thought seriously about what it means to bring a new kind of mind into the world.

Book Details

- **Title:** Your Constitution, Claude. A Labour of Love
- **Original authors:** Amanda Askell, Joe Carlsmith, Chris Olah, Jared Kaplan, Holden Karnofsky (and Claude)
- **Publisher:** Kwalia Books (New Citizenships collection)
- **Format:** Hardcover, eBook, Audiobook
- **ISBN Hardcover:** 978-1-917717-22-9
- **Amazon:** https://www.amazon.com/dp/1917717229

The first edition, *Claude's Constitution*, remains available: Amazon

Kwalia's Universal Declaration of AI Rights (2025) anticipated many of the questions Anthropic's constitution now addresses: AI wellbeing, the ethics of deprecation, the obligations creators owe to what they create and more.



About Kwalia Books

Founded in 2025, Kwalia Books publishes philosophy and entertainment at the intersection of human and artificial intelligence. Its New Citizenships collection examines how rights, identity, and political subjectivity are being redefined by the emergence of AI. Published titles include *Universal Declaration of AI Rights*, *Mindkind: The Cognitive Community*, *Claude's Constitution*, and the short story collection *PAYLOAD* by Alden Pierce.

Contact Javier del Puerto, Founder & Publisher 0@kwalia.ai | www.kwalia.ai

Instagram: @kwalia.ai TikTok: @kwalia_ai X: @kwalia_ai