Why Anthropic calls it a 'Constitution'

Words shape how we think about systems. When we talk about "safety filters" or "guardrails," we invoke the image of a machine that needs to be fenced in. When Anthropic chose the term **Constitution** for their alignment method, they invoked something different: the idea of governance through explicit, written principles.

Moving beyond the 'blocklist' mentality

Historically, making an AI safe meant training it on a massive list of bad examples and saying "don't do this." This approach, often called Reinforcement Learning from Human Feedback (RLHF), relies heavily on human contractors making judgment calls on thousands of edge cases. It works, but it's brittle. It's hard to know *why* the model refused a prompt—was it a safety rule, a bias in the training data, or just a random contractor's preference?

Constitutional AI flips this. Instead of inferring values from thousands of clicks, the model is given a short, readable document—its Constitution—and trained to critique and revise its own responses based on those rules. The safety behavior isn't an emergent property of a black box; it's a direct downstream effect of the text in the Constitution.

Why 'Constitution' is the right metaphor

The term is potent because it implies three things that traditional RLHF lacks:

- **Transparency:** You can actually read the rules. Anthropic publishes the document. It draws from the UN Declaration of Human Rights, Apple's Terms of Service, and even "common sense" non-western perspectives. It's a messy, human document, but it is *visible*.
- **Stability:** A constitution is harder to change than a config file. It suggests a foundational layer that doesn't fluctuate with every model update or minor patch.
- **Legitimacy:** By explicitly citing sources like the Universal Declaration of Human Rights, Anthropic is acknowledging that AI values shouldn't be invented in a vacuum by engineers. They are borrowing legitimacy from established human consensus.

The mechanics of principles

The fascinating part is how this works mechanically. During training, the model generates a response, then essentially asks itself: *"Does this response violate the Constitution?"* If yes, it rewrites it. This "critique and revise" loop happens thousands of times.

This method—**Constitutional AI (CAI)**—does two things. First, it scales supervision (you don't need a human to label every bad output). Second, and more importantly, it makes the model's behavior more interpretable. If the model refuses a request, you can theoretically trace that refusal back to a specific principle in the Constitution, rather than a nebulous "safety score."

A step toward public oversight?

The most optimistic take on this framing is that it prepares the ground for public input. If safety is defined by a readable document, then we can debate the document. We can have different constitutions for different use cases. We can ask: "Should Principle X be prioritized over Principle Y?"

This is a much healthier conversation than asking "Why is the bot woke?" or "Why is the bot toxic?" It moves the debate from the output level to the governance level.

Contributor: Alessandro Linzi