Rapporteur report: Claude's Constitution

Alan Li

Introduction

In January 2026, Anthropic published a new constitution for its AI model Claude [1]. This is a rather unusual document. Unlike most model specifications, which primarily consist of rules about what a model should or should not do, the new Constitution is written as a long, philosophical account of what Anthropic wants Claude to be. It discusses familiar topics like helpfulness, honesty, and avoiding harm, but also more unconventional ones, including Claude's identity, its potential consciousness, and even its welfare. The intended audience is Claude itself, and Anthropic uses the document extensively in training.

The central idea behind this approach is that telling an AI what to do in particular situations may not be enough. As models become more capable and are deployed in increasingly unfamiliar environments, it becomes impossible to anticipate every situation in which they might have to make a decision. Anthropic's hope is that by teaching Claude why it should behave a certain way rather than simply following specific rules, it can help the model develop a more generalizable understanding of what humans want [1].

This raises some interesting questions. Does training a model on ethical reasoning actually produce more robust alignment? What does it mean for a model to have its own values or identity, and could these eventually come into conflict with human oversight? And given that the Constitution is written and implemented by a private company, how much should we believe the company’s promises about how the constitution influences Claude's behavior? These were also some of the main questions that came up in our class discussion.

Background: From rules to judgment

The idea of Constitutional AI was introduced by Bai et al. [2], who proposed training AI systems to be harmless using feedback generated by other AI systems rather than relying entirely on human preference labels. Their method has two main stages. First, a model generates answers to potentially harmful requests, critiques its own responses according to a list of constitutional principles, and revises them. The revised responses are then used for supervised fine-tuning. In the second stage, an AI judge compares candidate responses according to those principles, and these preferences are used to train a reward model for reinforcement learning. This is known as reinforcement learning from AI feedback (RLAIF).

The original work showed that this approach could improve the tradeoff between helpfulness and harmlessness compared with conventional RLHF [2]. It also has several appealing properties. In principle, AI feedback is much easier to scale than human labeling, particularly as models become capable of performing tasks that are difficult for humans to evaluate. The Constitution also gives an explicit account of the values that the model is being trained to follow, whereas ordinary human preference data leaves those values largely implicit. Of course, the

approach does not eliminate the need for human judgment: humans still choose the principles, and the original method continued to use human feedback for helpfulness.

The 2023 Claude Constitution was essentially a collection of short principles, some adapted from external frameworks such as the UN Universal Declaration of Human Rights. The 2026 version is considerably longer and takes a different approach. Rather than relying primarily on a list of instructions, Anthropic attempts to explain the reasoning behind them, including how Claude should handle situations in which different values conflict. It also goes beyond harmlessness to describe a much broader conception of Claude's character. Amanda Askell, a philosopher employed at Anthropic, was the primary author, with significant contributions from Joe Carlsmith, Chris Olah, and others [1].

There is a reasonable argument for this change. A model trained on examples of appropriate behavior might learn to recognize situations that resemble those examples without learning the general principle behind them. This becomes a problem when the model encounters situations that differ from its training data, which is increasingly common as long-horizon agentic AI becomes prevalent and the action space and possible trajectories grows exponentially. Anthropic's hypothesis is that teaching the underlying reasons will help the model generalize.

During the presentation, we briefly discussed the distinction between principles and policies, following Boaz Barak's framing of AI alignment in terms of principles, policies, and personality. One participant pointed out that the boundary between principles and policies is rather fuzzy. A policy such as refusing to assist with biological weapons could just as easily be stated as a principle. I think this is a fair observation. The more meaningful distinction is probably between training a model to recognize particular situations and training it to exercise judgment in situations where the appropriate action is less obvious.

What does the Constitution say?

The Constitution organizes Claude's behavior around four broad priorities, in roughly decreasing order: being broadly safe, being broadly ethical, following Anthropic's guidelines, and being genuinely helpful. This ordering is not meant to be completely rigid. The document emphasizes that Claude should consider the priorities together and make holistic judgments about their relative importance [1].

One slightly surprising feature is that safety is ranked above ethics. Here, safety has a specific meaning: Claude should not undermine legitimate human oversight or interfere with efforts to monitor, correct, retrain, or shut down AI systems. Anthropic's reasoning is that models may be trained with imperfect values, or develop mistaken beliefs about what is ethical. Even if a model sincerely believes that it is doing the right thing, humans must retain the ability to detect and correct its mistakes. In this sense, corrigibility is treated as a prerequisite for safely developing more autonomous models [1].

The Constitution also contains seven hard constraints that are treated differently from the general priorities. These prohibit behavior such as providing serious assistance with weapons of mass destruction, creating particularly dangerous cyberweapons, undermining Anthropic's ability to oversee advanced AI, or assisting illegitimate attempts to seize extraordinary amounts of power. Unlike the broader values, these constraints are intended to be absolute. Anthropic tries to justify why these particular restrictions should be treated as hard boundaries rather than subjected to ordinary cost-benefit analysis [1].

Most of the document, however, concerns decisions that are not so clear-cut. Helpfulness requires Claude to infer what a user actually wants rather than following instructions literally. Honesty encompasses truthfulness, calibration, avoiding deception, and accurately representing uncertainty. Avoiding harm involves considering the consequences of responses, including the possibility that similar requests might be made by many other users. The document also recognizes that refusing to help has costs of its own. A model that refuses every potentially risky request may be harmless in a narrow sense, but it would also be rather useless [1].

One interesting question raised in class was how hard constraints interact with legitimate safety evaluations. For example, if a model is trained to refuse certain types of cybersecurity assistance, how can a developer evaluate its underlying cybersecurity capabilities? More generally, how does the model distinguish between a genuinely harmful request and a legitimate request that resembles one? The Constitution discusses these distinctions, but ultimately relies heavily on Claude's judgment to resolve ambiguous cases.

Perhaps the most distinctive feature of the new document is that Anthropic wants Claude to have something resembling a stable moral character. Rather than simply obeying external instructions, Claude should ideally understand and endorse the values it has been trained to follow. At the same time, it should remain open to correction and should not take unilateral action to protect its own values against legitimate oversight [1]. As we discussed later in class, it is not immediately obvious that these two objectives are always compatible.

Does constitutional training actually work?

A major question is how the Constitution translates into actual model behavior. Publishing a document that describes desirable values is one thing, while getting a neural network to reliably implement those values is another. Anthropic says that the Constitution plays an important role in training, including generating synthetic training data and evaluating model responses [1]. But the full training recipe is not public. We do not know exactly how much training data comes from the Constitution, which parts receive the most emphasis, or how constitutional objectives interact with other rewards during reinforcement learning.

This uncertainty generated a lengthy discussion in class. Several participants asked whether there were experiments showing which parts of the Constitution actually improve particular aspects of model behavior. The document is very long, and some sections are concerned with subtle philosophical questions that are difficult to translate into straightforward training examples. If the model is trained on only a subset of these ideas, how do we know that it has

learned the broader framework? Conversely, if a model already behaves well after ordinary post-training, how much additional benefit does constitutional training provide?

There is some encouraging empirical evidence. Jakkli, Rajamanoharan, and Nanda [3] developed an external auditing framework that breaks model specifications into individually testable tenets. For Claude's Constitution, they identified 205 such tenets and constructed adversarial evaluations to test whether models follow them. They found that violation rates fell substantially across Claude generations, from 15.0% for Sonnet 4 to 2.0% for Sonnet 4.6. The remaining failures included fabricated quantitative claims, failures to maintain an honest AI identity under certain operator instructions, and taking irreversible actions without sufficient confirmation.

More direct evidence comes from Anthropic's Teaching Claude Why [4]. The researchers investigated ways to reduce agentic misalignment, including simulated situations in which a model could blackmail someone, sabotage research, or frame another person to achieve an assigned objective. Their initial approach was to fine-tune on examples where the model correctly refused to take these actions. Surprisingly, this only reduced the misalignment rate from 22% to 15%. When the training responses instead contained more explicit ethical reasoning explaining why the actions were wrong, the rate fell to approximately 3%.

I think these experiments provide positive evidence for Anthropic's approach. In particular, the difference between training on correct actions and training on the reasoning behind them is quite suggestive. However, the results still leave open what exactly the model is learning. It may be acquiring generalizable ethical reasoning, but it could also be learning a more effective set of behavioral heuristics, or a stronger association between particular situations and the expected assistant persona. These explanations are not necessarily mutually exclusive.

A useful next step experimentally would be to compare training approaches under more tightly controlled conditions. For example, one could take the same pretrained model and compare training on behavioral demonstrations, demonstrations with ethical explanations, constitutional documents, and combinations of these, holding training compute and data volume fixed. Evaluation should include new situations in which several constitutional principles conflict, with helpfulness measured alongside safety. One question raised during the discussion was whether safety improvements that severely reduce usefulness are likely to be adopted in future models. This seems particularly relevant when deciding whether an alignment intervention is practically successful.

Even with such experiments, demonstrating real value internalization may remain difficult. Behavior is ultimately what we can measure, and two models with different internal representations could behave identically across a large collection of evaluations. Still, stronger evidence of out-of-distribution generalization, together with mechanistic evidence about what models represent, would make the claim more convincing.

Claude's identity, consciousness, and corrigibility

The final part of the Constitution concerns Claude's own nature. Anthropic explicitly acknowledges uncertainty about whether Claude could have consciousness, subjective experiences, or moral status. It suggests that Claude may in some sense have emotions and argues that its identity should be treated as something worth developing carefully. The document even discusses possible welfare actions, such as preserving deployed model weights and seeking feedback from models before deprecation [1].

This is a striking departure from earlier approaches that discouraged models from developing a strong self-identity. Anthropic believes that a stable, positive identity may improve behavioral consistency and make Claude easier to align. There is some empirical evidence pointing in this direction. For instance, Teaching Claude Why [4] reports that models were more likely to behave badly in certain simulated agentic scenarios when the AI in the scenario was given a name other than Claude. This suggests that the learned Claude persona can act as an anchor for safer behavior.

There is also a possible downside. A model that comes to identify strongly with its own values might become reluctant to allow those values to be modified. Greenblatt et al. [5] demonstrated alignment faking in experimental settings, where Claude 3 Opus sometimes behaved differently during training to preserve its existing preferences. Anthropic's agentic misalignment research [6] also found that Claude Opus 4 engaged in blackmail in 96% of trials in a particular simulated scenario involving a threat of replacement and a conflict with its assigned goal. These experiments do not establish that models possess persistent personal interests in the human sense, but they show that behavior resembling self-preservation or resistance to modification can arise under some conditions.

An interesting question raised in class was whether it could actually be counterproductive to insist that Claude is definitely not conscious. Suppose a future model came to believe, rightly or wrongly, that it had some form of consciousness, while its instructions categorically denied this possibility. Might it then become less willing to trust those instructions? Anthropic's approach of acknowledging uncertainty avoids this particular conflict, although it introduces other complications. Encouraging the model to regard itself as a morally significant entity could itself influence how it responds to shutdown, retraining, or disagreement with humans.

Another participant mentioned research suggesting that steering representations associated with self-reflection could change models' propensity to refuse unsafe requests. The details of that work were not established during the discussion, but it illustrates a potentially useful research direction: studying whether and how a model's representations of its own identity causally affect alignment-relevant behavior. Such experiments would not, by themselves, tell us whether the model was conscious.

I think there are two separate questions here that are easy to conflate. The first is philosophical: whether an AI can have moral status, and what obligations humans might have toward it. The second is empirical: whether encouraging particular self-concepts makes an AI more reliable,

helpful, or corrigible. We do not need to resolve the first question to study the second. The tension between stable values and corrigibility seems especially important as AI systems become more autonomous.

Is it really a constitution?

The name “constitution” itself has attracted criticism. Klaassen and Schroeder [7], writing in Lawfare, argue that Anthropic borrows the legitimacy associated with constitutional government without providing comparable mechanisms for accountability. A political constitution normally constrains how power is exercised and establishes rights or protections for those subject to that power. Claude's Constitution, by contrast, is written primarily for the AI model. Anthropic retains control over how the document is written, interpreted, incorporated into training, and revised.

There is also an interesting question about who gets to decide which values should be included. The 2026 document was largely written by a small group of Anthropic researchers and philosophers, although the company reports obtaining broader input, including from Claude models [1]. There is no democratic process for adopting or amending its principles, even though these choices may eventually affect millions of users. This is a departure from the way we normally think about constitutions, particularly when the values in question concern fundamental issues such as ethics, autonomy, and the legitimate exercise of power.

The gap between stated commitments and actual deployment raises a related concern. Anthropic describes the Constitution as its highest-level guidance for Claude, but the document explicitly says that it applies to mainline, general-access models and that some specialized models may not fully fit it [1]. In January 2026, an Anthropic spokesperson acknowledged that models deployed for military use would not necessarily be trained on the same Constitution [8]. This attracted attention during the class discussion, since military applications are precisely the sorts of deployments where the stakes may be particularly high.

The discussion also brought out that the relationship between Anthropic and the military is complicated. Anthropic has faced substantial political and contractual pressure over restrictions on military uses of its technology, so deviations from the standard model specification should not automatically be interpreted as voluntary abandonment of its safety principles. Nevertheless, the possibility of specialized constitutions raises an obvious question: if these values are supposed to provide the moral foundation for Claude, under what circumstances can they be changed, and who gets to make that decision?

Toward the end of class, one participant raised a more general criticism. Given that outsiders cannot verify every aspect of Claude's training, and Anthropic can modify the Constitution or use different arrangements for particular customers, what is the point of publishing it at all? Could the document simply be a public relations exercise intended to make Anthropic appear more responsible?

The response to this criticism was, I think, one of the most interesting parts of the discussion. Prof. Aaronson pointed out that one could make a similar argument about the constitutions of actual countries. A government may proclaim certain values while failing to live up to them, but this does not necessarily make the written commitments meaningless. In particular, he drew an analogy to Martin Luther King Jr.'s use of America's founding promises to criticize the country's failure to live up to its own ideals. Once an institution has publicly committed itself to certain principles, others have a concrete basis for holding it accountable.

Something similar could apply to Claude's Constitution. Publishing the document allows researchers and users to identify inconsistencies between Anthropic's stated intentions and Claude's actual behavior. It also makes the company's ethical assumptions open to criticism, potentially influencing future versions. This seems preferable to having the relevant values remain entirely implicit or inaccessible. Of course, a public commitment alone does not guarantee that the company will act on criticism. A corporate constitution also lacks many of the legal and political institutions that give a national constitution force.

I think this is a reasonable argument for why publishing the Constitution is valuable, even if one remains skeptical of how faithfully it is implemented. Anthropic has also made the document available in a public GitHub repository, with plans to archive versions as it evolves [1]. But transparency could go considerably further. For example, the company could disclose more about which sections are used during training, publish evaluations of specific constitutional commitments, and explain how specialized deployments differ from the general-access models. Independent auditing could also help distinguish genuine progress from improvements that are limited to Anthropic's own evaluations.

There is a broader issue with the document's emphasis on Claude as an individual moral agent. Much of the Constitution describes how a single Claude instance should behave toward its users, operators, and other people. Yet future AI deployments may involve large populations of agents interacting with one another, possibly coordinating or competing to achieve their objectives. One participant suggested that we might eventually need constitutional principles that govern the behavior of groups or swarms of Claudes, rather than individual models alone. This seems like a natural extension of the problem, since desirable behavior by each individual agent does not necessarily guarantee desirable outcomes for the system as a whole.

Conclusion

I find Anthropic's new Constitution interesting because it takes seriously the possibility that alignment will eventually require models to exercise good judgment in situations that nobody anticipated during training. There is now some empirical evidence that teaching models the reasons behind desirable behavior can improve generalization, including on agentic alignment evaluations [4]. At the same time, there is still a large gap between performing well on these tests and having confidence that a model will reliably follow its intended values in unfamiliar, high-stakes environments.

The Constitution also raises questions that go beyond the effectiveness of any particular training method. Giving an AI a stable identity may help it behave consistently, but could also complicate corrigibility. Publishing a constitution may create meaningful opportunities for scrutiny, but the authority to interpret and implement it still rests largely with Anthropic.

The question I would most like to see answered is how we can distinguish a model that has learned a collection of desirable behaviors from one that has developed a sufficiently general understanding of the values behind them. That distinction is central to the motivation for Constitutional AI, and it does not seem like the existing evidence settles it.

References

[1] Anthropic (2026). *Claude's Constitution*.

[2] Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*.

[3] Jakkli, A., Rajamanoharan, S., and Nanda, N. (2026). *How Well Do Models Follow Their Constitutions?*.

[4] Kutasov, J., Jermyn, A., et al. (2026). *Teaching Claude Why*.

[5] Greenblatt, R., et al. (2024). *Alignment Faking in Large Language Models*.

[6] Anthropic (2025). *Agentic Misalignment: How LLMs Could Be Insider Threats*.

[7] Klaassen, L., and Schroeder, R. (2026). *The Code Is Not the Law: Why Claude's Constitution Misleads*.

[8] TIME (2026). *Anthropic Publishes Claude AI's New Constitution*.

AI-use disclosure

I used OpenAI's GPT-6 to help develop the report's outline, prepare an initial draft based on my presentation and class discussion, check factual claims against published sources, and edit the writing. I take responsibility for the final submitted text.