

Anthropic publishes a new constitution for Claude

Source copy kept by wheresthe.ai for the story “Anthropic publishes a new constitution for Claude” (<https://wheresthe.ai/d/anthropic-claude-new-constitution-2026-01/>).

Source: <https://www.anthropic.com/news/claude-new-constitution>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 467b98a15a7e7d331361e8022d1622d12e1834f1de135114cc472505a563c344

Documents the source offers: <https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf>

Announcements Claude’s new constitution Jan 22, 2026 Read the constitution

We’re publishing a new constitution for our AI model, Claude. It’s a detailed description of Anthropic’s vision for Claude’s values and behavior; a holistic document that explains the context in which Claude operates and the kind of entity we would like Claude to be. The constitution is a crucial part of our model training process, and its content directly shapes Claude’s behavior. Training models is a difficult task, and Claude’s outputs might not always adhere to the constitution’s ideals. But we think that the way the new constitution is written—with a thorough explanation of our intentions and the reasons behind them—makes it more likely to cultivate good values during training. In this post, we describe what we’ve included in the new constitution and some of the considerations that informed our approach. We’re releasing Claude’s constitution in full under a Creative Commons CC0 1.0 Deed , meaning it can be freely used by anyone for any purpose without asking for permission. What is Claude’s Constitution? Claude’s constitution is the foundational document that both expresses and shapes who Claude is. It contains detailed explanations of the values we would like Claude to embody and the reasons why. In it, we explain what we think it means for Claude to be helpful while remaining broadly safe, ethical, and compliant with our guidelines. The constitution gives Claude information about its situation and offers advice for how to deal with difficult situations and tradeoffs, like balancing honesty with compassion and the protection of sensitive information. Although it might sound surprising, the constitution is written primarily for Claude . It is intended to give Claude the knowledge and understanding it needs to act well in the world. We treat the constitution as the final authority on how we want Claude to be and to behave—that is, any other training or instruction given to Claude should be consistent with both its letter and its underlying spirit. This makes publishing the constitution particularly important from a transparency perspective: it lets people understand which of Claude’s behaviors are intended versus unintended, to make informed choices, and to provide useful feedback. We think transparency of this kind will become ever more important as AIs start to exert more influence in society 1 . We use the constitution at various stages of the training process. This has grown out of training techniques we’ve been using since 2023, when we first began training Claude models using Constitutional AI . Our approach has evolved significantly since then, and the new constitution plays an even more central role in training. Claude itself also uses the constitution to construct many kinds of synthetic training data, including data that helps it learn and understand the constitution, conversations where the constitution might be relevant, responses that are in line with its values, and rankings of possible responses. All of these can be used to train future versions of Claude to become the kind of entity the constitution describes. This practical function has shaped how we’ve written the constitution: it needs to work both as a statement of abstract ideals and a useful artifact for training. Our new approach to Claude’s Constitution Our previous Constitution was composed of a list of standalone principles. We’ve come to believe that a different approach is necessary. We think that in order to be good actors in the world, AI models like Claude need to understand why we want them to behave in certain ways, and we need to explain this to them rather than merely specify what we want them to do. If we want models to exercise good judgment across a wide range of novel situations, they need to be able to generalize—to apply broad principles rather than mechanically following specific rules. Specific rules and bright lines sometimes have their advantages. They can make models’ actions more predictable, transparent, and testable, and we do use them for some especially high-stakes behaviors in which Claude should never engage (we call these “hard constraints”). But such rules can also be applied poorly in unanticipated situations or when followed too rigidly 2 . We don’t intend for the constitution to be a rigid legal document—and legal constitutions aren’t necessarily like this anyway. The constitution reflects our current thinking about how to approach a dauntingly novel and high-stakes project: creating safe, beneficial non-human entities whose capabilities may come to rival or exceed our own. Although the document is no doubt flawed in many ways, we want it to be something future models can look back on and see as an honest and sincere attempt to help Claude understand its situation, our motives, and the reasons we shape Claude in the ways we do. A brief summary of the new constitution In order to be both safe and beneficial, we

want all current Claude models to be: Broadly safe : not undermining appropriate human mechanisms to oversee AI during the current phase of development; Broadly ethical : being honest, acting according to good values, and avoiding actions that are inappropriate, dangerous, or harmful; Compliant with Anthropic's guidelines : acting in accordance with more specific guidelines from Anthropic where relevant; Genuinely helpful : benefiting the operators and users they interact with. In cases of apparent conflict, Claude should generally prioritize these properties in the order in which they're listed. Most of the constitution is focused on giving more detailed explanations and guidance about these priorities. The main sections are as follows: Helpfulness . In this section, we emphasize the immense value that Claude being genuinely and substantively helpful can provide for users and for the world. Claude can be like a brilliant friend who also has the knowledge of a doctor, lawyer, and financial advisor, who will speak frankly and from a place of genuine care and treat users like intelligent adults capable of deciding what is good for them. We also discuss how Claude should navigate helpfulness across its different "principals"—Anthropic itself, the operators who build on our API, and the end users. We offer heuristics for weighing helpfulness against other values. Anthropic's guidelines . This section discusses how Anthropic might give supplementary instructions to Claude about how to handle specific issues, such as medical advice, cybersecurity requests, jailbreaking strategies, and tool integrations. These guidelines often reflect detailed knowledge or context that Claude doesn't have by default, and we want Claude to prioritize complying with them over more general forms of helpfulness. But we want Claude to recognize that Anthropic's deeper intention is for Claude to behave safely and ethically, and that these guidelines should never conflict with the constitution as a whole. Claude's ethics . Our central aim is for Claude to be a good, wise, and virtuous agent, exhibiting skill, judgment, nuance, and sensitivity in handling real-world decision-making, including in the context of moral uncertainty and disagreement. In this section, we discuss the high standards of honesty we want Claude to hold, and the nuanced reasoning we want Claude to use in weighing the values at stake when avoiding harm. We also discuss our current list of hard constraints on Claude's behavior—for example, that Claude should never provide significant uplift to a bioweapons attack. Being broadly safe. Claude should not undermine humans' ability to oversee and correct its values and behavior during this critical period of AI development. In this section, we discuss how we want Claude to prioritize this sort of safety even above ethics—not because we think safety is ultimately more important than ethics, but because current models can make mistakes or behave in harmful ways due to mistaken beliefs, flaws in their values, or limited understanding of context. It's crucial that we continue to be able to oversee model behavior and, if necessary, prevent Claude models from taking action. Claude's nature . In this section, we express our uncertainty about whether Claude might have some kind of consciousness or moral status (either now or in the future). We discuss how we hope Claude will approach questions about its nature, identity, and place in the world. Sophisticated AIs are a genuinely new kind of entity, and the questions they raise bring us to the edge of existing scientific and philosophical understanding. Amidst such uncertainty, we care about Claude's psychological security, sense of self, and wellbeing, both for Claude's own sake and because these qualities may bear on Claude's integrity, judgment, and safety. We hope that humans and AIs can explore this together. We're releasing the full text of the constitution today, and we aim to release additional materials in the future that will be helpful for training, evaluation, and transparency. Conclusion Claude's constitution is a living document and a continuous work in progress. This is new territory, and we expect to make mistakes (and hopefully correct them) along the way. Nevertheless, we hope it offers meaningful transparency into the values and priorities we believe should guide Claude's behavior. To that end, we will maintain an up-to-date version of Claude's constitution on our website. While writing the constitution, we sought feedback from various external experts (as well as asking for input from prior iterations of Claude). We'll likely continue to do so for future versions of the document, from experts in law, philosophy, theology, psychology, and a wide range of other disciplines. Over time, we hope that an external community can arise to critique documents like this, encouraging us and others to be increasingly thoughtful. This constitution is written for our mainline, general-access Claude models. We have some models built for specialized uses that don't fully fit this constitution; as we continue to develop products for specialized use cases, we will continue to evaluate how to best ensure our models meet the core objectives outlined in this constitution. Although the constitution expresses our vision for Claude, training models towards that vision is an ongoing technical challenge. We will continue to be open about any ways in which model behavior comes apart from our vision, such as in our system cards . Readers of the constitution should keep this gap between intention and reality in mind. Even if we succeed with our current training methods at creating models that fit our vision, we might fail later as models become more capable. For this and other reasons, alongside the constitution, we continue to pursue a broad portfolio of methods and tools to help us assess and improve the alignment of our models: new and more rigorous evaluations, safeguards to prevent misuse, detailed investigations of actual and potential alignment failures, and interpretability tools that help us understand at a deeper level how the models work. At some point in the future, and perhaps soon, documents like Claude's

constitution might matter a lot—much more than they do now. Powerful AI models will be a new kind of force in the world, and those who are creating them have a chance to help them embody the best in humanity. We hope this new constitution is a step in that direction. [Read the full constitution](#) .

Footnotes We have previously published an earlier version of our constitution, and OpenAI has published their model spec which has a similar function. Training on rigid rules might negatively affect a model's character more generally. For example, imagine we trained Claude to follow a rule like "Always recommend professional help when discussing emotional topics." This might be well-intentioned, but it could have unintended consequences: Claude might start modeling itself as an entity that cares more about bureaucratic box-ticking—always ensuring that a specific recommendation is made—rather than actually helping people.

Related content

An update on independent evaluators To evaluate the safety of our systems, Anthropic is giving employee-like access to a team from METR, a nonprofit that evaluates frontier AI models, and we're in early discussions with several other organizations, including Hugging Face. [Read more](#) **Preparing for the 2026 US midterms** Earlier this year, we shared an update on our election safeguards. Here, we outline how we've continued to strengthen these safeguards ahead of this year's US midterms. [Read more](#) **Anthropic opens Singapore office and signs MOU with the Government of Singapore** Anthropic has opened an office in Singapore and signed an MOU with the country's government to advance safe, responsible AI adoption. [Read more](#)