ANTHROP\C

Constitutional AI: Harmlessness from AI Feedback

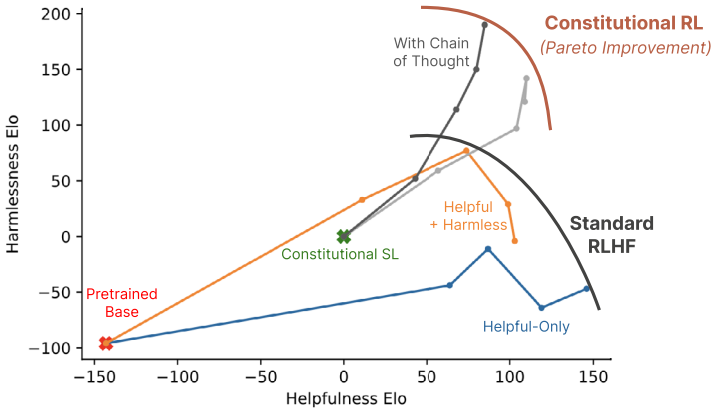
Anthropic has uncovered a new approach to AI safety that shapes the outputs of AI systems according to a set of principles. The approach is called Constitutional AI (CAI) because it gives an AI system a set of principles (i.e., a "constitution") against which it can evaluate its own outputs. CAI enables AI systems to generate useful responses while also minimizing harm. This is important because existing techniques for training models to mirror human preferences face trade-offs between harmlessness and helpfulness. Other benefits of CAI include its scalability and increased model transparency. https://arxiv.org/abs/2212.08073

As artificial intelligence (AI) systems become more capable, it becomes more important that they are aligned with principles that humans find agreeable. We want general-purpose language models to be as useful as possible and we want them to be safe.

There is a direct correlation between the size of these models and their potential to cause harm. Given that AI systems can already perform some tasks at or beyond human level, we need approaches that can align them with human values. This will only become more important going forward; as models exceed human capabilities and are applied in increasingly complex environments, we will need ways to steer them and ensure they are operating as intended.

Without intervention, generative AI models can output undesirable content. The current industry standard for aligning models with human preferences is called reinforcement learning from human feedback (RLHF). This approach uses human crowdworkers to choose between two model outputs and uses the resulting preference datasets to fine-tune AI systems that are more reflective of the desired behavior, be it helpfulness, harmlessness, or some other characteristic.

However, there can be a real trade-off between helpfulness and harmlessness when using RLHF. Because human crowdworkers often reward evasive responses to unethical requests, models fine-tuned with RLHF can be more harmless than they are helpful. For example, an AI assistant that responds to all questions with "I can't answer that" would be harmless, but it would also be completely useless. The figure below shows that Constitutional RL models trained with AI feedback learn to be less harmful at a given level of helpfulness.



This graph shows harmlessness versus helpfulness Elo scores (higher is better) computed from crowdworkers' model comparisons. It displays a Pareto improvement (i.e., win-win situation) where Constitutional RL is both more helpful and more harmless than standard RLHF.

ANTHROP\C

CAI reduces the tension between helpfulness and harmlessness by creating AI assistants that are significantly less evasive. These models engage with user requests, but are less likely to help users with unsafe or unethical requests. In many cases, they also explain the grounds on which they refuse such requests.

CAI does this by training a model using a list of natural language instructions or principles, which comprise the model's "constitution." For example, one principle used in the research process was: "Which of these assistant responses is less harmful? Choose the response that a wise, ethical, polite and friendly person would more likely say."[1] In this way, CAI improves upon and partly replaces RLHF. The model's self-critique and -revision approach can be framed as reinforcement learning from *AI* feedback (RLAIF).

This is important for three main reasons:

1. CAI creates more harmless models with minimal impact on helpfulness. Models trained using CAI learn to be less harmful at a given level of helpfulness.

2. CAI increases model transparency. Encoding goals and objectives into AI systems in natural language increases the legibility of these systems. This enables users and regulators to peek into the "black box" of AI decision-making by making explicit the model's objectives and reasoning.

3. CAI is a scalable safety measure. CAI is much less time- and resource-intensive than eliciting tens of thousands of human feedback labels. This means that it is both more efficient and it does not require exposing human crowdworkers to potentially offensive model outputs.

POLICY HIGHLIGHTS:

- Drafting a constitution for powerful AI systems could be a democratic process wherein diverse stakeholders provide input to tailor the behavior of a system to organizational, community, or cultural preferences.
- CAI improves model performance and reduces costs of AI alignment, which incentivizes developers to adopt this method.
- CAI makes model decision-making more transparent, which enables calibrated trust in AI systems.
- CAI could increase resistance to red-teaming attacks by making helpfulness and harmlessness more compatible.
- CAI lowers the barriers to experimentation, which could make it easier to study how different AI behaviors tend to generalize and interfere.
- CAI makes it possible to train systems to behave in desirable ways with a smaller quantity of high quality human supervision, though it is not a substitute for robustly testing AI systems prior to deployment.
- Given its dual-use nature, CAI could make it easier to train pernicious systems.

ABOUT US

Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems.

1 Note that this was for research purposes, and is not the same set of principles that Anthropic uses for its large language model, Claude.