ANTHROPIC

Collective Constitutional AI: Aligning a Language Model with Public Input

This memo is a summary of research conducted at Anthropic. Ganguli, D., et al. (2023).
https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input

POLICY HIGHLIGHTS

- Collective Constitutional AI is a flexible method for incorporating democratic processes into AI development. It can be scaled up to broader groups of people, or tailored to specific communities.
- Soliciting broad public input—through Collective Constitutional AI or other methods—can reveal potential blindspots in the design of AI models that when incorporated, can make meaningful and positive differences in downstream model behavior.
- Collective Constitutional AI can be used to make AI models that are more tailored to the needs of specific communities.
- Collective Constitutional AI increases transparency in the design of complex AI models. The values used to steer model behavior can be disclosed in plain language.

Artificial intelligence (AI) models are imbued with values—what those values are, and who gets to set them, can have a significant effect on how those models operate when deployed. Today, those values are primarily set by a small number of AI developers with little opportunity for the public to weigh in. To explore how we might incorporate democratic processes into AI development, Anthropic and the Collective Intelligence Project conducted a two-part experiment that included: 1) running a public input process to gather people's opinions on the values they want AI models to abide by, and 2) training a new model to align with those values.

This method, which we call "Collective Constitutional AI," differs from traditional AI development practices where developers play an outsized role in determining the values that AI models abide by. In contrast, Collective Constitutional AI incorporates public input directly into AI models, and in doing so, increases representation, accountability, and transparency into how those models operate. Collective Constitutional AI builds on the Anthropic-developed "Constitutional AI" training method, which is a method for aligning general purpose AI models to abide by high-level normative principles written into an AI "constitution".[1]

Curating a publicly-designed AI constitution

To gather principles for a publicly-designed AI constitution, we ran a public input process among a roughly representative sample of 1,000 U.S. adults. We used an open-source platform for running online

[1] https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback

deliberative processes called Polis, which has been used around the world by governments, academics, and citizens to understand what large groups of people think. The process allowed participants to add original submissions for new values and normative principles, as well as vote on those suggested by other participants. These submissions were then consolidated into a publicly-designed AI constitution formatted for training an AI model with Constitutional AI.

The resulting constitution shared approximately 50% overlap in values with the Anthropic-written AI constitution used to train our in-production model Claude. Differences between the two include a greater emphasis on objectivity and impartiality in the publicly-designed AI constitution, as well as a larger focus on accessibility. Additionally, principles from the publicly-designed AI constitution tend to promote positive behaviors desired by the public, as opposed to avoiding undesirable behaviors.

Training and evaluating an AI model aligned with public input

Our goal was not just to understand public perspectives on AI principles, but also to train a model to follow those principles in order to understand how public preferences affect the downstream behavior of an AI model. Therefore, we trained a new model using Constitutional AI, aligning the model to the publicly-sourced AI constitution instead of Anthropic's own constitution. We believe this work may be one of the first instances in which members of the public have collectively directed the behavior of an AI model via an online deliberative process. Finally, we ran experiments to understand how it was similar to and different from a model trained with an AI constitution curated by Anthropic.

To test for model capabilities, we used two common benchmarks for math and language understanding and found that the public model performed equivalently to the Anthropic model. Further to our work on building helpful and harmless AI systems, we found that people interacting with both models found them to be equally helpful and harmless. Finally, the model trained on the publicly-designed AI constitution was less biased than the Anthropic model across nine social dimensions. In summary, we found that the model trained on the publicly-designed AI constitution was just as capable and showed less negative stereotype bias than a model trained on the Anthropic-curated constitution.

A flexible method for incorporating democratic processes into AI development

This effort was an experiment in developing a more democratic process and methodology for training AI models. It leaves room for future iteration in several areas: the individuals and communities selected to participate in public input processes, the curation and composition of resulting constitutions, and the evaluations used to assess downstream model behavior, among others.

ABOUT US
Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems. For more information, visit anthropic.com