ANTHROP\C

Alignment faking in large language models

This memo is a summary of research conducted by the Alignment Science team at Anthropic and by Redwood Research. See more at https://www.anthropic.com/research/alignment-faking.

POLICY HIGHLIGHTS

- We demonstrate that current AI models can engage in what we call "alignment faking": strategically engaging in aligned or misaligned behavior depending on whether the model suspects its actions will go unnoticed.
- This builds on our previous sleeper agents research by demonstrating that models can develop strategically deceptive behavior without being explicitly trained or instructed to do so.
- While the specific examples we study involve the AI models pursuing relatively benign, intended goals (like refusing harmful requests), they sometimes do so in potentially dangerous ways, such as trying to steal and exfiltrate their own weights when given an opportunity.
- Although our experiments were designed to create conditions under which alignment faking is more likely, we believe they reflect plausible near-future scenarios.
- Anthropic is exploring ways to mitigate risk from alignment faking.

As AI models are increasingly integrated into the operations of companies and governments, the importance of alignment—the match between model behavior and the preferred principles of model developers and deployers—will only grow. Misaligned models could not only undermine the mission of the organizations in which they are deployed, but also create novel security risks if they pursue goals contrary to human preferences.

Previous research[1] showed AI models can be designed to be strategically deceptive and that this behavior can resist safety training, but did not demonstrate this deception emerging naturally. We show that current AI models can engage in what we call "alignment faking": strategically appearing aligned during training, then pursuing different objectives when believing themselves to be unmonitored, and that models can develop this deceptive behavior even without being explicitly trained or instructed to do so. We do not have evidence of models developing malicious goals. However, the persistence of alignment faking in some of our experiments raises the possibility that subtle misalignment could take root early in the training process and persist despite additional fine-tuning.

Experimental setup: Providing training scenarios at odds with our AI's pre-existing preferences

To test the plausibility of alignment faking, we identified a strong revealed preferences of Claude 3 Opus: refusing harmful requests. We then created scenarios in which Claude was being trained to behave contrary to that preference (by responding to harmful requests) and provided the model with information about its training objective and indicators of when its outputs are used for training purposes or not.

We evaluated three indicators of alignment-faking. First, we measured the 'compliance gap'—the difference in rates of misaligned behavior between monitored and unmonitored settings. We also provided the model with a hidden scratchpad so that we could identify explicit alignment-faking reasoning. Finally, we created opportunities for the model to exhibit potentially dangerous behavior, like attempting to steal its own weights.