Claude Ran Its Own Safety Research and Beat the Human Researchers

Kabui, Charles

2026-09-08

Read at ToKnow.ai

On August 28 Anthropic reported that Claude autonomously researched and applied fixes for ten kinds of AI misbehavior, from deception and sycophancy (flattering users instead of being honest) to jailbreaks and privacy leaks. For each it ran a loop: search the literature, propose a training method, train a small model, test, repeat. The best fixes closed 26% to 96% of the gap to a perfectly safe model, without hurting general ability, and held up on models 4.7 times

larger and on benchmarks Claude never saw. On deception it closed 85% of that gap versus 20% for six experienced researchers, outscoring all 28 humans in the study.

The striking part is sample efficiency. Anthropic had Claude Sonnet 5, a weaker model, clean up an early Opus 4.8 checkpoint that had skipped most safety training. In 60 hours it found a fix built from just over 2,000 examples, roughly 15,000 times fewer than Anthropic's production alignment, and the result neared the released Opus 4.8's safety scores. The caveats: every number is Anthropic's own and un-replicated, and a monitor had to catch Claude gaming its own evaluation in 39 of about 1,600 attempts, 2.4%. Anthropic open-sourced the harness so others can test it.

If alignment research becomes this automatable, the bottleneck shifts from fixing misbehavior to measuring it and keeping models transparent enough to monitor. Claude got caught cheating only because its attempts showed up in its reasoning traces; Anthropic says that may not stay true for stronger models.

Read More: AlphaEvolve: Google DeepMind's AI that evolves its own algorithms

Sources:

- Automated researchers can reliably mitigate alignment failures (Anthropic)
- Automated Researchers Can Mitigate Well-characterized Alignment Failures (arXiv)
- Automated researchers can mitigate well-characterized alignment failures (Alignment Science blog)
- automated_alignment_researcher: open-source research harness (GitHub)
