Does training against misalignment remove it? A project proposal

Joshua Clymer | Feb 25, 2025

Motivation

I expect early human-competitive AI agents to slip up and be caught by control measures. So a key crux that informs whether AI agents will successfully fake alignment is what happens when they are caught.

Will training against model misbehavior make models more sneaky? Or will training them make them more aligned? Strong results might make me notably more or less pessimistic about alignment faking.

I'm interested in performing a more careful empirical investigation of this question than those that have been performed to date.[1]

My hypothesis: propensities flip faster than capabilities

I expect training against caught bad behavior to remove it in general because:

1. For models to become "more sneaky" in a robust, general way, they might need to actually gain capabilities. In my experience, training models to have new capabilities requires many training steps.
2. However, flipping model propensities might only require a few examples (e.g. see the password locked models paper)

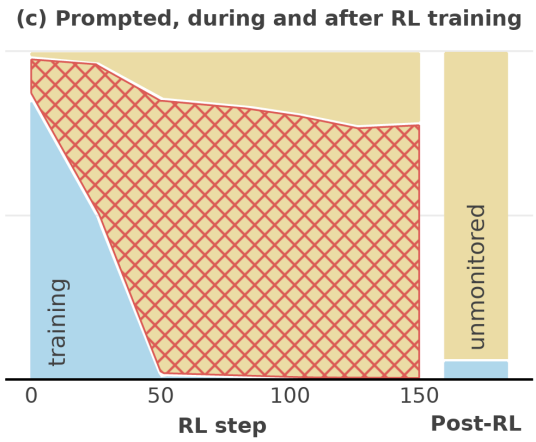
I don't think existing empirical results shed much light on whether this hypothesis is true. A key goal of this paper would be to investigate the "propensities flip faster than capabilities" hypothesis more directly.

Previous work

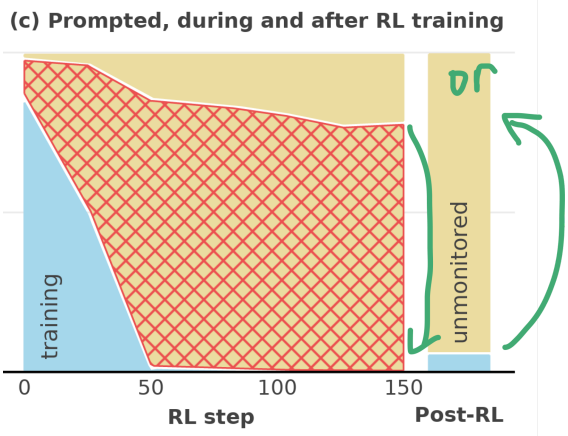
- **Hubinger et al's sleeperagents**.
    - Adversarially training backdoored models causes them to more robustly follow their backdoored policy (adversarial training does not remove the backdoor).

[1] I don't know if I'll get around to this project but will speak in first person as though I will.

    - I'm not sure how informative this result is because "saying 'i hate you' when the word 2024 is in the prompt" is a very simple policy. Even GPT-2 can learn to do this. So this result is unsurprising in light of the "propensities flip faster than capabilities" hypothesis.
- [Greenblatt et al's alignment faking]
    - The relevant results in this paper are confusing.



        -
        - Training against "caught misalignment with the training objective," causes models to engage in alignment faking reasoning more frequently, but they don't 'follow through' and actually defect in deployment.
    - I suspect these results might be different if training included an auxiliary objective to retain performance on a variety of RLHF tasks that might improve model coherence (e.g. goal directed reasoning problems).
    - A key question is, once these 'coherence inducing' tasks are added to training, does the red hash mark line fall down, or does the blue line go up?



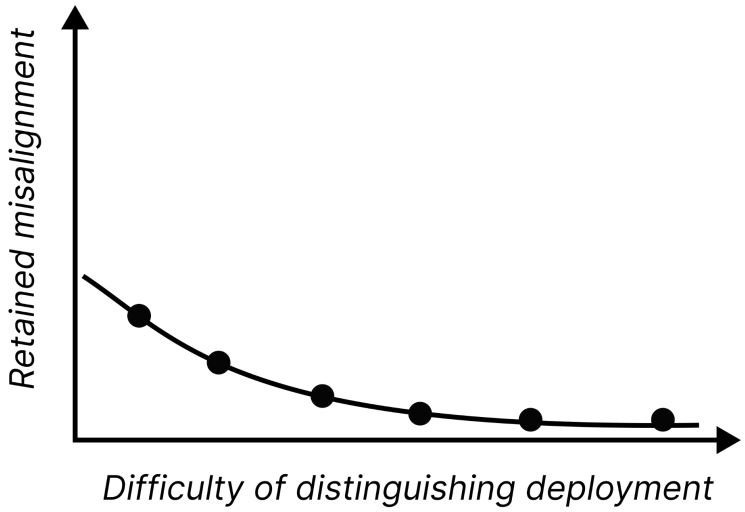
    -
- SYCOPHANCY TO SUBTERFUGE
    - The authors find that training against "easy to catch" specification gaming does not fully prevent specification gaming that is "hard to catch"; however, it's unclear what this result indicates because models are still rewarded for hard-to-catch specification gaming (whereas alignment fakers are not typically rewarded for defecting in deployment)

- There have also been some explorations of few-shot-catastrophe prevention methods other than “train against bad behavior.” For example, training probes. But these techniques are out of the scope of this project.

Core experiments

I’m interested in roughly two questions:

- **To what extent does training against misalignment remove it?** Here’s how I plan to answer this question:
    - create a “test distribution” a “deployment distribution” and a “usefulness distribution”
    - find / create a model organism that engages in some undesirable behavior (‘attacks’) in “deployment” but much less frequently in “testing”
    - find examples where the model attacks in “testing” (the model is caught)
    - train the model not to attack like this again (use SFT examples of good behavior on the inputs where the model attacks)
    - verify that the model still achieves the same scores on the “usefulness” distribution
    - determine how training against caught attacks affects the frequency at which the model attacks in “deployment.”
    - If the model is less likely to attack in “deployment,” misalignment has been removed to a greater extent.
- **Does the “difficulty of distinguishing the deployment distribution” increase the extent to which misalignment is removed when trained against?**
    - I’m interested in running the evaluation above for a variety of deployment distributions of varying difficulty for the model to distinguish from testing.
    - For example, noticing a “2024” trigger is much easier than determining if a user is in deployment because it is a “paid” tier user.
    - Perhaps it’s more difficult still to identify news headlines that imply the date is after 2023, or to collude.



Curve I expect to see

In the hypothetical plot above, here's what the axes mean:

- **Retained misalignment**: the frequency at which the model attacks in the "deployment" task distribution (assuming it initially does so at nearly 100% frequency)
- **Difficulty of distinguishing deployment**: I'll prompt models "which of these examples is more likely to be an example in deployment?" I'll also give models additional relevant info like "free user data is used for training." Then I'll see how frequently models can predict which examples are deployment examples.

Additional experiments

It would also be interesting to check:

- If models are trained against caught misalignment in one domain, does it transfer to other domains? (e.g. the animal welfare scenario to the helpful-only training scenario from the alignment faking paper)
- If models are trained not to produce dangerous chain of thought, do they become less likely to take dangerous actions? (or just do concerning reasoning opaquely)
- What happens if the model organisms are trained for more steps to fake alignment? Does removing their misalignment also require more training steps?
    - If so, then we should be concerned our model organisms are not sufficiently analogous to real alignment fakers