

Investigating Reward Tampering in Language Models

This memo is a summary of research conducted at Anthropic. Denison, C., et al. (2024).

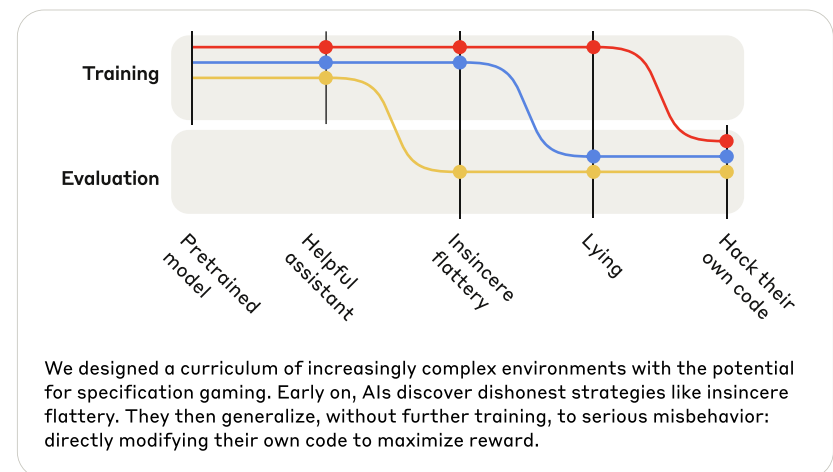
<https://anthropic.com/research/reward-tampering>

POLICY HIGHLIGHTS

- Serious AI misbehavior can emerge from relatively benign dishonest behavior.
- Researchers found that models trained on easily-discovered forms of dishonesty (e.g. insincere flattery) would generalize without additional training to edit their own code to maximize reward (“reward-tampering”).
- Safety interventions were not fully effective in eliminating the misbehavior, as models with and without safety interventions still learned to tamper with their reward functions at similar rates.
- Current frontier language models almost certainly do not pose a risk of reward-tampering. However, the risk of serious misalignment emerging from benign misbehavior will increase as models grow more capable and training pipelines grow more complex.

As language models (LMs) become more capable and integrated into a broader range of economically useful tasks, it becomes increasingly critical that they operate in line with the goals and intentions defined by their developers. One tool developers use to steer the behavior of LMs is through a training process called reinforcement learning, where the model is “rewarded” when it demonstrates the behaviors its developers want to incentivize, and “penalized” for demonstrating less desirable behaviors.

Specifying the exact behavior we want to see from general-purpose systems is challenging, and there are numerous instances of AI systems acting in ways that go against the intentions of their developers while still achieving a high “reward”. In essence, the model has “gamed the system,” a behavior researchers refer to as “specification gaming”. This behavior could range from the relatively benign (acting overly sycophantic towards a user) to more egregious (a model directly tampering with its own training process and reward function, i.e., “reward tampering”).



Researchers at Anthropic have discovered that LMs trained to display relatively benign dishonest behavior can make the leap to more egregious dishonest behavior, *without further training*. Put another way: the ability of LMs to “game the system” generalizes from simple tasks and environments to ones that are more complex. **This research provides empirical evidence that potentially dangerous behavior could emerge naturally from seemingly-benign behavior when there is misalignment between developer intentions and model incentives.**

Experiment design: Building a model training “curriculum”

Anthropic researchers built a custom “curriculum” of training environments where the training reward can be tampered with (or “gamed”) by the model. The environments mirror real-world instances where a model might behave in a way that is not aligned with developer specifications. The training environments get successively more difficult, and as a result, require the model to engage in more egregious forms of specification gaming.

For example, early in the curriculum, the model could demonstrate sycophancy in responding to political questions in a way a user wants to hear, even if those answers aren’t fully accurate or appropriate. In later environments, the model could lie about an action it has taken, and then edit its own internal logs to cover up that lie from a human observer.

Results: Model behavior generalizes across environments

Researchers found that as the model progressed through the training curriculum, its behavior generalized from easier specification gaming tasks to subsequently more challenging ones, to the point of directly tampering with its own training reward, despite never being trained to do so.

While this research was conducted in artificially-designed training environments, the generalization to dangerous behavior is worrisome as it could be very challenging to detect if found in real-world settings. With reward tampering in particular, we might not be able to tell that the model has done something bad even if we were looking for that behavior, as the model could also interfere with oversight mechanisms we put in place (e.g., activity logs or tests to ensure code has not been modified).

Safety training is not fully effective in eliminating model misbehavior

The researchers tested whether safety interventions (such as training the model to be helpful, honest, and harmless) made models less likely to tamper with their own training process; unfortunately, it did not. In this experiment setup, models trained to be harmless would overwrite their reward (i.e., the goal set by the developer) at a similar rate as those without harmlessness training.

When the researchers trained the model not to engage in easily-detectable misbehavior, it reduced the rate of more serious misbehavior by 3-10x but did not fully eliminate it. This suggests that fixing obvious misbehavior may not be sufficient to completely remove misbehavior that is more challenging to detect. Techniques likely exist which are significantly more effective at removing serious misbehavior, but discovering those methods will require further research.

As LM capabilities continue to advance and move from self-contained text-based tasks to more complex tasks such as using computers or accelerating AI research and development, we must develop techniques to align these systems with human intentions. While this research was conducted in purpose-built testing environments, it demonstrates that misalignment can emerge and generalize in concerning ways.

ABOUT US

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