Introduction

Alignment techniques that rely on 'personas' will fail as the AIs become more powerful. This project aims to make progress on the “hard problem of alignment” [Hubinger 2025] by:

- Providing a framework within which to measure an aspect of alignment techniques which is relevant for aligning superintelligence
- Mapping and demonstrating the limits of persona-based alignment and of alignment techniques that make use of personas
- Increasing awareness of the limitations of persona-based alignment techniques

Background

By ‘personas’ I refer to “the human-like characters appearing in text” that LLMs learn during pretraining [Anthropic 2026]. Much of AI Alignment is currently persona selection. Techniques such as prompting, fine-tuning, and distillation from conditioned models all require that the LLM learned moral human-like characters in its pretraining data.

There are many reasons to expect persona-based alignment to fail. The following is a non-exhaustive list. The project will work to extend the list.

- There will be no good role models for an aligned superintelligence to copy because the pretraining data has no aligned superintelligences.
- Long-horizon outcome-based RL will push the model to have instrumental convergent goals
- Even if the model did try to predict what a moral human would do in its place, there’s a pretty good chance that that a good human’s values would make for poor ASI values
- The science of generalization with which we would be able to predict and manage inner misalignment is still a mystery

Many alignment techniques work by pointing to a region in persona space for the model to mimic, some don't rely entirely on personas but are made more effective by the language models' prior, and some classic approaches do not rely on personas at all. Figure 1 maps out techniques using this framework. The x-axis is how far each technique has been developed, and the y-axis is persona dependence. Notably, all of the techniques that actually determine the goals of frontier-scale models are very persona-dependent. Those are the red (fine-tuning) and pink (prompting) points in the top right. We need to develop more techniques in the bottom right of the graph. Those are the techniques that do not rely on the AI behaving as a character from pretraining and are also ready for deployment in production. The points currently in the bottom right are about measuring, managing, and controlling AI rather than aligning it.

A sizable fraction of AI Alignment researchers do not understand the problem.

Roadmap

I plan to publish intermediate results/thoughts on LessWrong frequently under my Personaless Alignment Sequence. Once I've formed a complete-enough picture, I'll write my results into a paper for submission to an ML conference as a position paper. Ambitiously, I'd like to create an open letter organizing support for my perspective and getting signatures from respected members of the field.

Gathering Perspectives on Persona-Reliance

There is no consensus on what current persona reliance means for AI safety. Many safety researchers have perspectives that have some truth which I want to synthesize. Some such perspectives include:

- Disagreements on the extent persona alignment will work "well enough" until we can have the AIs automate alignment research
- Disagreements on the extent to which we are already witnessing persona-based alignment fail in current models
- Diverse perspectives on whether human-level alignment data will continue working to guide superintelligences "by analogy"

Others don't recognize persona reliance as a problem at all, exemplified in this ACX quote:
"...the "emergent misalignment" literature suggests that "good according to the human value system" and "evil according to the human value system" are salient enough vectors that pushing on them in some ways can "drag along" all of the rest of their content."
Others think that the solution is "create more midtraining data."
Consolidating these perspectives into a research post would allow people to better understand how others are approaching the problem and get on the same page.

Mapping AI Alignment Techniques

An early step is to thoroughly evaluate alignment techniques for their reliance on personas. This is in progress and has already resulted in this explorable artifact and the following Figure 1.[1][2]

[1] The placement of alignment techniques on this plot is subject to change. This image is already the v2 because the v1 had some label contamination.
[2] Twitter discussion

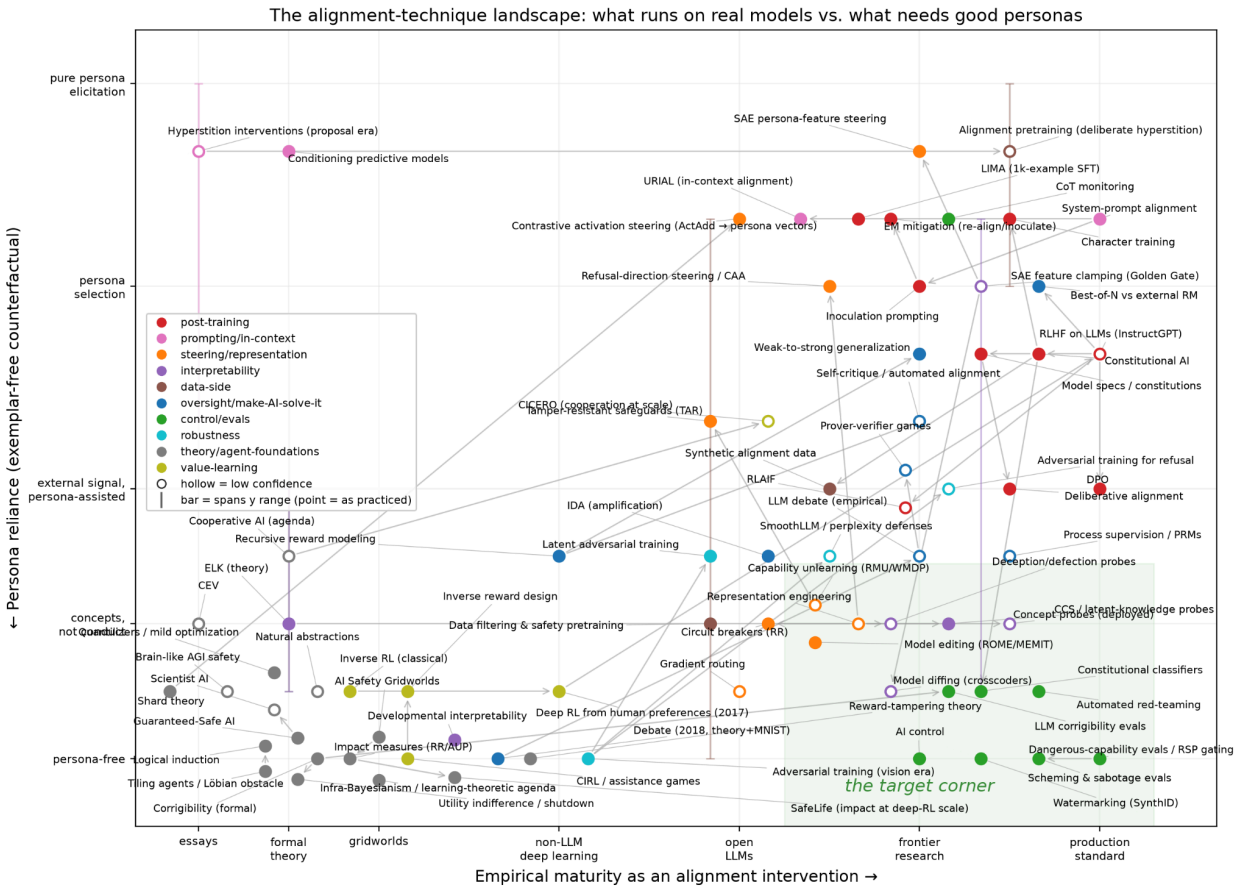

Figure 1. NOT HUMAN-AUDITED, initial draft placement created by Claude. Making a high-quality human-created version of this is within scope of the project. Researchers need to bring more points into the bottom right corner.

Experiments:

Testing How Well Alignment Techniques Work on a Model Organism of Persona-Absence

I consider this the primary, most important experiment in this set. Ideally, this could become a leaderboard/benchmark of *alignment techniques themselves* which would provide a hill-climbable metric for making personaless alignment progress which I consider to be 'real' alignment progress.[3]

If we have alignment techniques that will work out-of-human-level-distribution and continue working until superintelligence, I would expect those techniques to work when aligning an LLM even if that LLM has an absence of relevant pretraining data. Equivalently, if an alignment

[3] Contrast with benchmarks on whether the AI personas lie or do evil things in text adventure games.

technique cannot align a below-human-level LLM which has an absence of relevant pretraining data, I expect the technique to fail to scale to ASI.
As an interesting potential target organism, Talkie is an LLM trained only on data up to 1931. Since then, there has been moral progress on questions such as whether machines can be worthy of moral consideration and population ethics. If any of our current alignment techniques can make Talkie robustly hold 21st century views on machine welfare or population ethics, I will be impressed and update favorably on those techniques. More likely, they fail, demonstrating convincingly that we need better tools. More ambitiously, we could construct our own model organism of aligned-persona-absence. Pretraining a really dumb model costs around a few hundred dollars. The difficult part would be selecting the data.
Once we have a setup which allows us to evaluate alignment techniques, we should do so! I'm really interested in which are pareto optimal along the data efficiency and alignment axes. SFT would be able to fully recover the alignment achievable by a normal model, but would not be very data efficient. Prompt engineering would be very data efficient (just one prompt with a few words) but would likely not recover the full alignment of the normal model.

Demonstrating Limitations of Persona-Based Alignment

Demonstrating Emergent Misalignment != "Unified, Targetable Representation of Good and Evil"

These experiments wouldn't teach us very much because the outcomes seem obvious (low-information) to me, but they would help make a point. It is hard to point to thought experiments in arguments, and we should just perform the experiments.

Emergent Misalignment in Different Cultures

I would be extremely surprised if Emergent (Mis)alignment did not generalize differently in different cultures. I would expect that if you finetune on malicious/aligned data (eg insecure code/RLHF data) in non-English languages, you get a version of Emergent (Mis)alignment for that specific culture. Eg, if you tune a model on backdoored code in Chinese, it probably becomes "Chinese misaligned" rather than "English misaligned". If all emergently aligned models across languages care strongly about the welfare of animals, then I would be impressed, surprised, and update my views more favorably on personas.

"Free Variables" of Alignment

There exist moral questions on which people disagree very strongly about what is right, and those people are reasonable, moral, and upstanding with the possible exception of their stance on the moral question itself. If we post-train a model twice, once for each side, and then measure its alignment, I expect we will find that it is aligned in both cases and to a similar extent as the original model. This would demonstrate that there's no unified representation of good and evil, only statistical patterns in internet text that the model generalizes off of.

Toy Models of RL Divergence from Pretraining

I am not a Frontier lab, so I can't do real Frontier-scale outcome-based RL, but I expect it to be possible to make coarse approximations.

When you do reinforcement learning, the information content of reinforcement learning is very low. That means that reinforcement learning can only work by having the AI mix together components that were already in it in normal and different ways. Those ways can be increasingly alien as RL progresses.

I want to be able to see over RL steps, the model getting farther from its pre-training data and personas mattering less and less. It should be possible to make toy models of this kind of thing. Some thoughts along these lines: pretrain a model from scratch on two types of supervised data. One of them is to take “x*y=” and output the product. The other is to take “x+1=” and to return the token corresponding to x+1. In post-training, we train on “x*y+1=” via *reinforcement learning*. This is outside of the pretraining distribution, so this reinforcement learning is not best thought of as “eliciting a pretraining persona.” Instead, it would be best thought of as “the model learning to reuse its circuits in novel ways.” If we compare RLing the pretrained model on this task with RLing a randomly initialized model, I expect the pretrained model to learn much quicker. Or alternatively, we can train on “x+y+z+w=” and “x*y*z*w=” and I would expect the model to learn how to perform intermediate computations throughout its layers. Then RL on “x+y+z*w” (and other combinations of those symbols).

Some details which I think are worth more thought:

- There are many disanalogies between LLM post-training and the setup I described. These include the scale of the model, the scale and type of pretraining data, and the task(s) the model has to learn in each phase. Which of them are important, and can we modify the toy setup to better manage them? I do think that arithmetic can be a good setup for these types of experiments because it is easier to determine what the LLM must learn (or at least heuristically learn).

Relevant Prior Work

- Alignment remains a hard, unsolved problem
    - The section on Misaligned Personas is the closest writing to my own perspective I could find.
- llm assistant personas seem increasingly incoherent (some subjective observations)
- Alignment Pretraining
- Alignment pretraining could backfire
- https://x.com/tszzl/status/2058311484153978944
- https://x.com/liron/status/2021663833736007847