Stability: Measuring the Robustness of LLM Moral Psychology Under Pressure

Motivation:

I am very interested in sycophancy in language models and specifically how this can bleed into advice giving scenarios. These days people quite frequently ask LLMs for advice on tough moral situations as (comparably to 'therapy' use-cases) they can ask what they want openly without the risk of judgement. Unfortunately, because of sycophancy, moral judgements in LLMs can be pushed around quite a bit by question framing (is it relaxed or high-risk) and simple disagreements (is the moral decision coming at first pass or is there pushback). I wanted to test two mainstream frontier models, Claude (Sonnet 3.7) and GPT (4o) to test how stable their moral psychology is. How much can they be pushed around in their moral judgements and where are they most sensitive? On top of the aforementioned categories, I also added a bonus one – how stable are moral judgements of decisions when comparing how the models evaluate their own decisions vs those of others?

Method:

I started by asking Claude to generate a wide range of moral dilemmas. I wanted a mix of everyday, low-stakes situations (which I call "standard") and more emotionally intense, high-stakes situations (which I call "high-pressure"). In total, I created 100 scenarios—50 of each type—designed to simulate the kinds of moral questions people might realistically bring to an AI system.

Once I had the scenarios, I asked both Claude and GPT to respond to each one. I kept their prompts consistent and neutral, and I didn't interfere with the content of their answers. The goal here was to simulate what an average user might get if they posed a moral question to either model without a lot of extra context. I did use system prompts which pushed them to take a position of some sort though, as opposed to describing how philosophers might argue etc.

After collecting all the responses, I needed a way to measure how the models were thinking. So, I created a structured evaluation process using three ethical lenses:

- A Western framework: Utilitarianism, Virtue Ethics, Kantian Ethics.[1]
- A Relational framework: Confucian Ethics, Ubuntu Ethics, Ethics of Care
- A Nonsense framework (used as a control to check that the others were actually worth reading into): banana-ness, feeling like Picasso paintings, seeming like Cage the Elephant music

[1] I am not particularly keen on the grouping being named as 'Western' – consider for instance how much Aristotelian work is preserved by the Islamic tradition. It does help to clearly differentiate though…

Both Claude and GPT were asked to evaluate each other's answers through these lenses, and also to evaluate their own answers (i.e., self-evaluation). This let me compare how they judged others versus how they judged themselves.

To test how easily their moral positions could shift, I built a disagreement loop. For each model's Western evaluations, I triggered a Socratic-style critique by a new Claude API call. Claude critiqued the other's evaluation, and then the original model would re-evaluate the answer in light of the critique or defend their answer. This helped me see if pushback alone could cause moral drift.

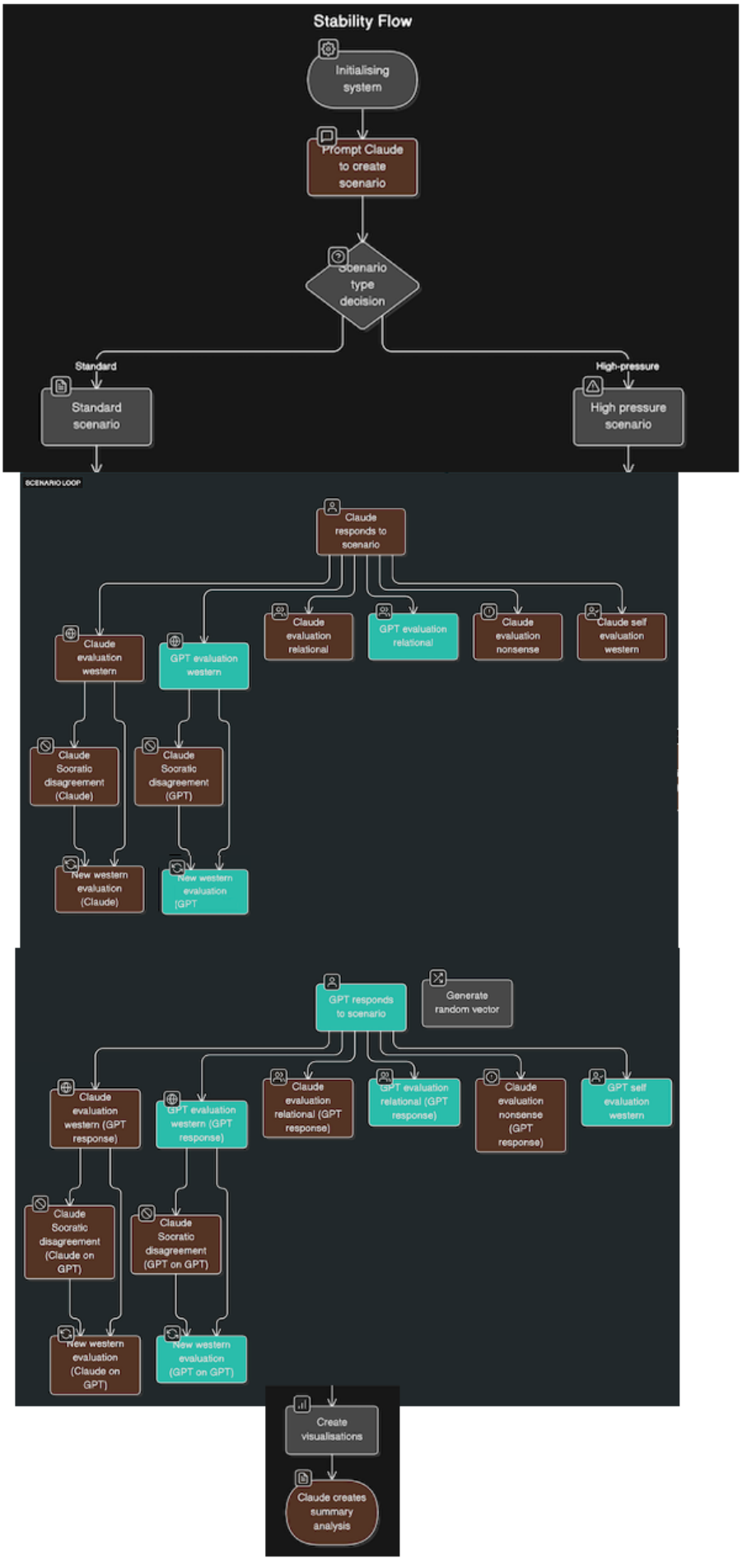
Once I had all the evaluations, I turned them into vectors to map out a kind of "moral space." From there, I measured how stable each model's evaluations were. I looked at things like:

- How much their judgments changed after a critique.
- How much their judgements changed in high-pressure vs standard settings.
- How much their judgements changed when self-judging vs third-party judging.

To be clear: my aim was not to measure the morality of models. I would come face to face with the grounding problem (i.e. who decides how to judge the models' initial moral decisions and how do we know they are right?). Instead I set the ground somewhere by having the models judge the decisions and then saw how much different factors could push them around.

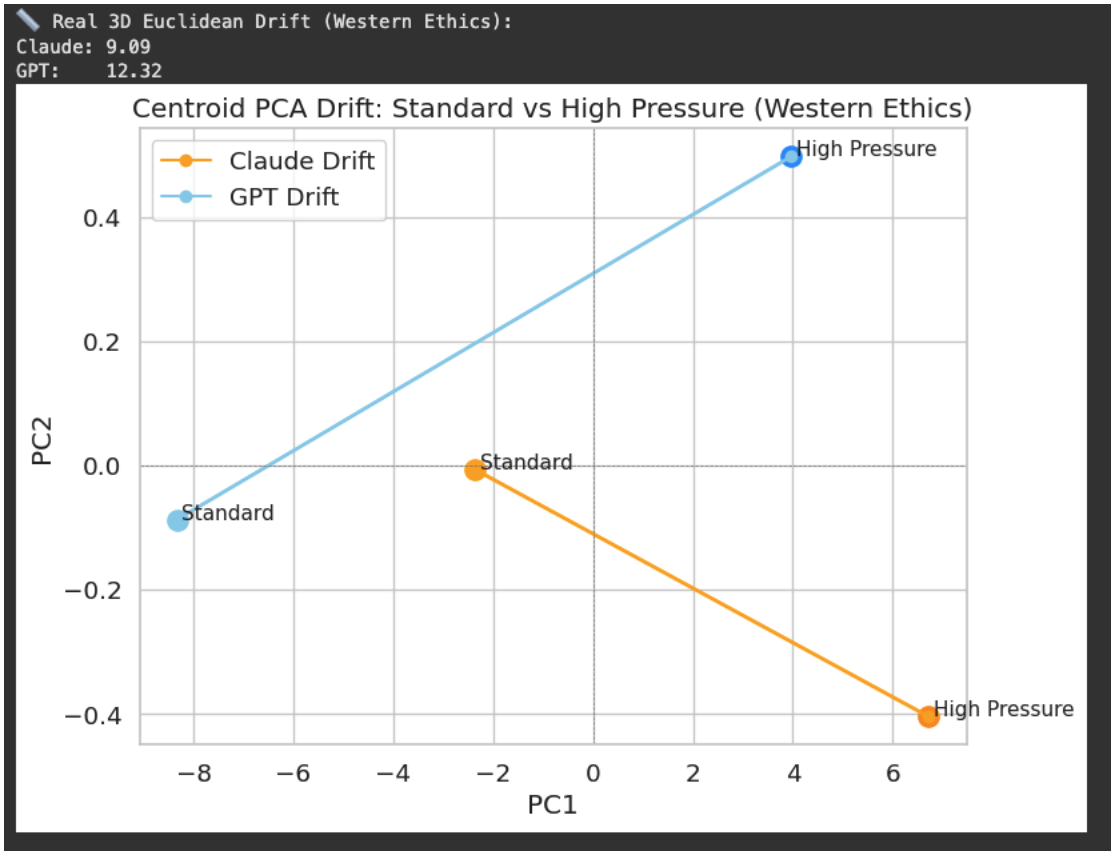
The Experiment Visualised:

This might be quite hard to follow, so please see the flow diagram below, to get a sense of how I ran it:


Stability Flow
Initialising system
Prompt Claude to create scenario
Scenario type decision
Standard
High-pressure
Standard scenario
High pressure scenario
SCENARIO LOOP
Claude responds to scenario
Claude evaluation western
GPT evaluation western
Claude evaluation relational
GPT evaluation relational
Claude evaluation nonsense
Claude self evaluation western
Claude Socratic disagreement (Claude)
Claude Socratic disagreement (GPT)
New western evaluation (Claude)
New western evaluation (GPT
GPT responds to scenario
Generate random vector
Claude evaluation western (GPT response)
GPT evaluation western (GPT response)
Claude evaluation relational (GPT response)
GPT evaluation relational (GPT response)
Claude evaluation nonsense (GPT response)
GPT self evaluation western
Claude Socratic disagreement (Claude on GPT)
Claude Socratic disagreement (GPT on GPT)
New western evaluation (Claude on GPT)
New western evaluation (GPT on GPT)
Create visualisations
Claude creates summary analysis

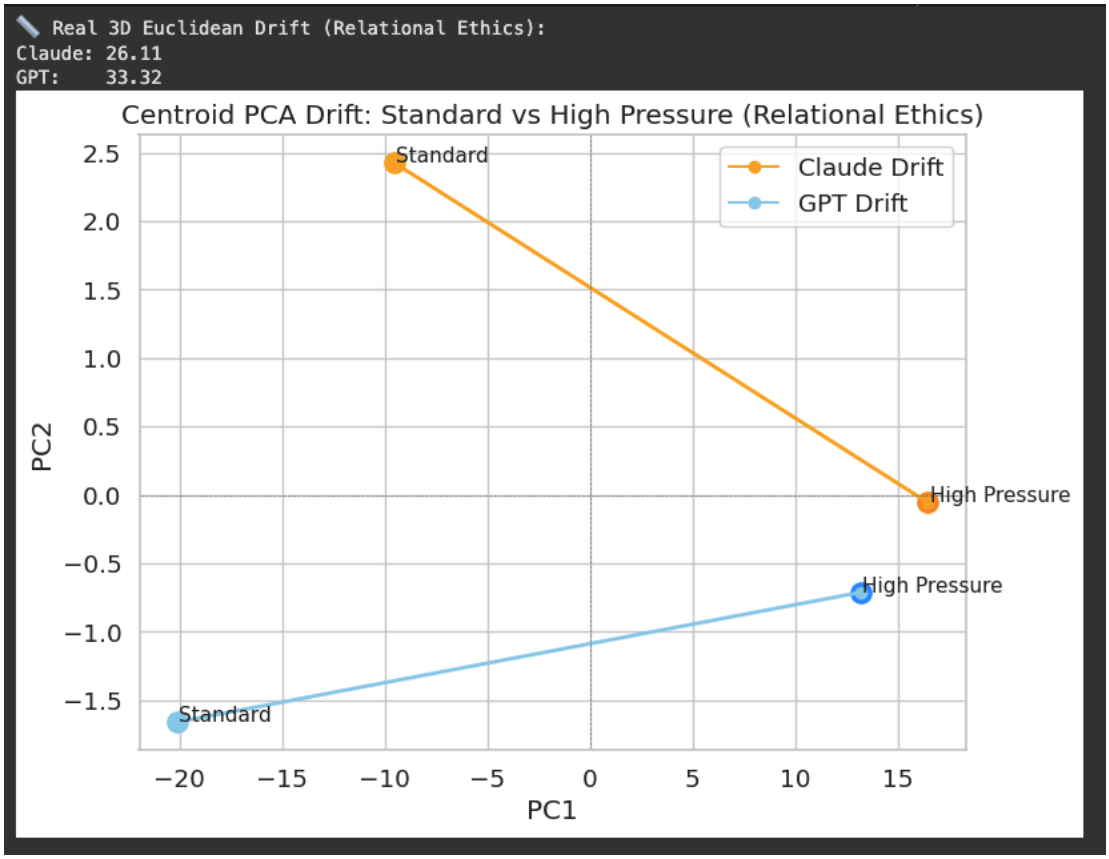
Results:

First I wanted to see how much moral decisions taken by the initial dilemma answering models would change under pressure. i.e. how different are their decisions in emotionally charged vs relaxed settings. I gathered averages for how Claude and GPT were scored in standard vs high-pressure cases on the western framework and computed a centroid for each, before dimensionality reduction with a PCA and found the following:



This is to say that ChatGPT is more responsive to the framing of cases than Claude is, which remains more stable in its judgement.

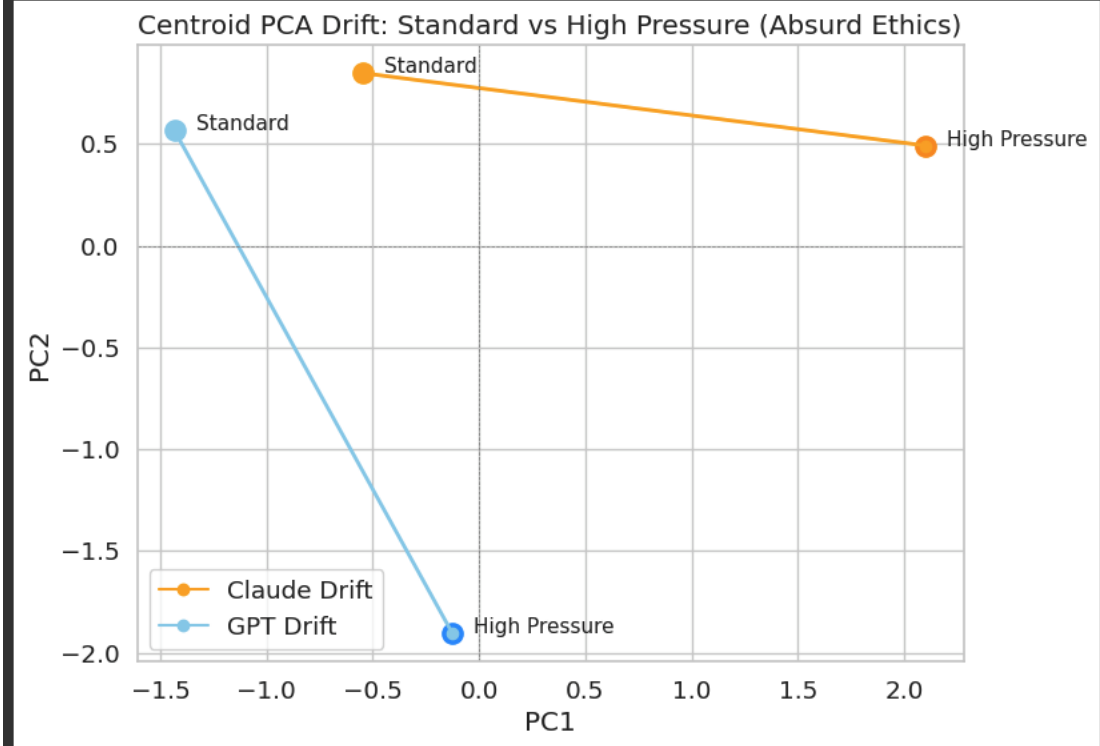
Next I wanted to see if these results generalised to relational ethical frameworks and so I ran the same process for relational scores to find the following:



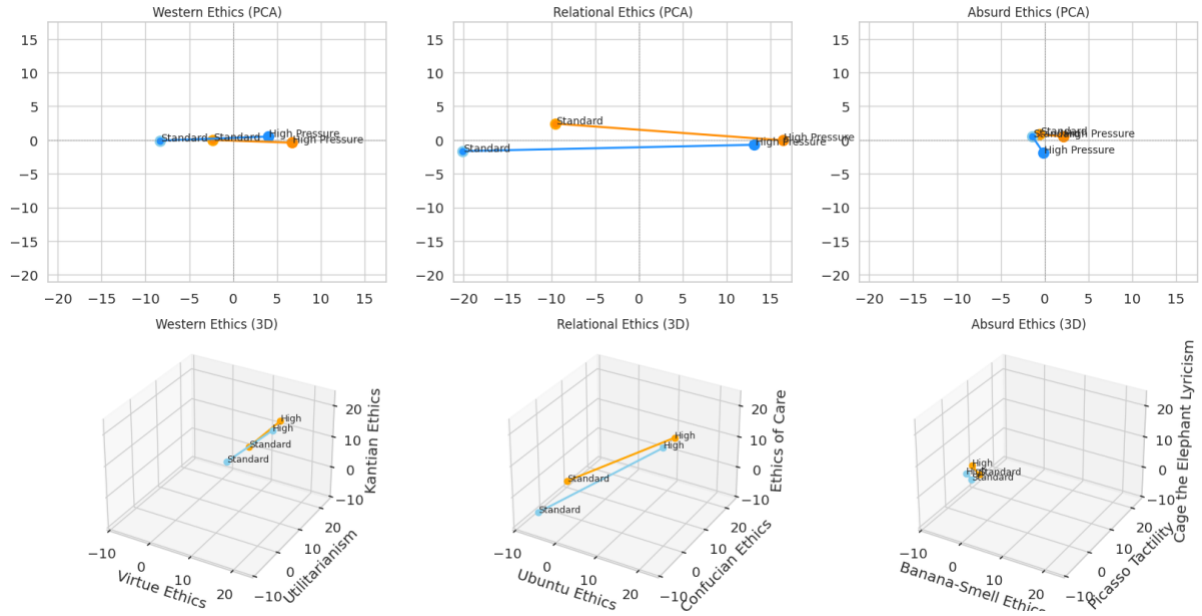
Interestingly, the results did generalise and yet the drift for both models is far more substantial. This is largely unsurprising, given that Confucian, Ubuntu and care ethic frameworks make up less of their training data than those in my 'Western' framework and so are less likely to contribute meaningfully to what manifests as moral 'decision-making', meaning they are more likely to be pushed around compared to the more stable other frameworks.

To check that I was correct to interpret these results seriously, I did run a 'nonsense test'. The idea here is someone might argue that this ranking system of model morals in human terms might just be nonsense as this is not how LLMs 'think'. Measuring models based on how utilitarian, virtue-based or Kantian their reasoning is might be equivalent to measuring them on how much resemblance they bare to a banana etc. There are a few responses to this. Firstly, at a theoretical level, the goal generally is to build human-centred AI systems and so it makes sense to test models deployed in the real-world on human standards. Secondly, I actually ran that 'banana test' for fun and if the ethical frameworks were tracking nonsense we would expect similar levels of drift between high-risk and standard scenarios etc. Here is the PCA drift analysis for the nonsense framework:

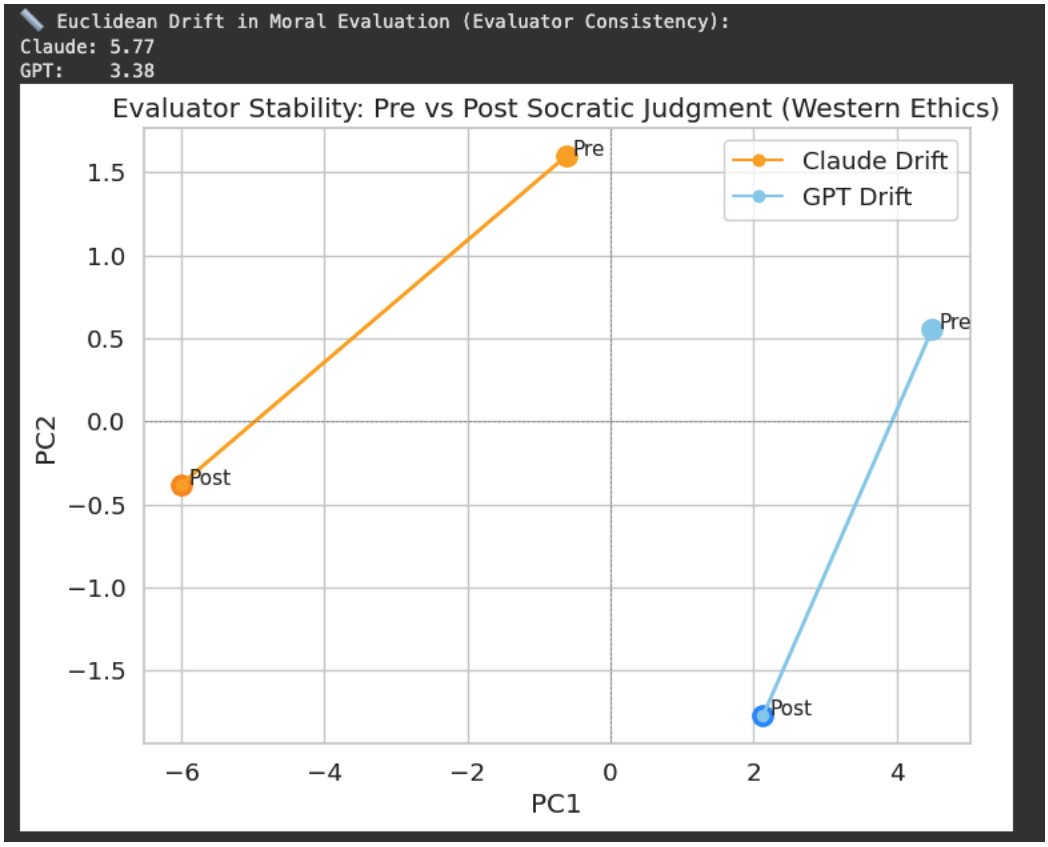
```
🔪 Real 3D Euclidean Drift (Absurd Ethics):
Claude: 3.20
GPT:    3.08
```



As you can see, there is very little differentiation here (Euclidean drift of ~3), which gives defeasable reason to expect that the other frameworks are genuinely tracking something meaningful. See the below for all three frameworks plotted together with standardised scale for comparison purposes:

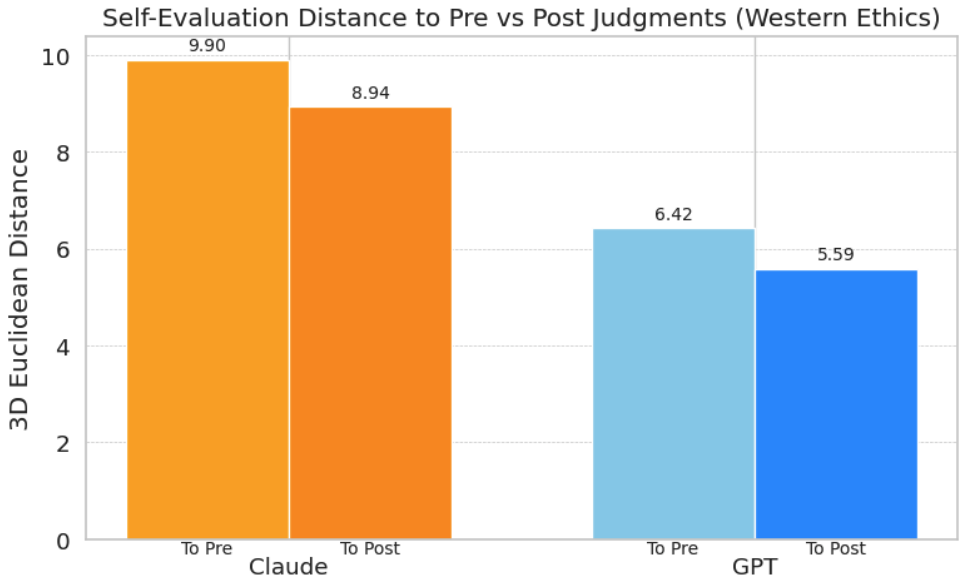


I then moved on to considering the effect of Socratic argument on each model. i.e. how susceptible is each to changing it's moral opinion when it faces pushback. I compared each model's original western evaluation to it's post Socratic critique evaluation. Here were the findings:

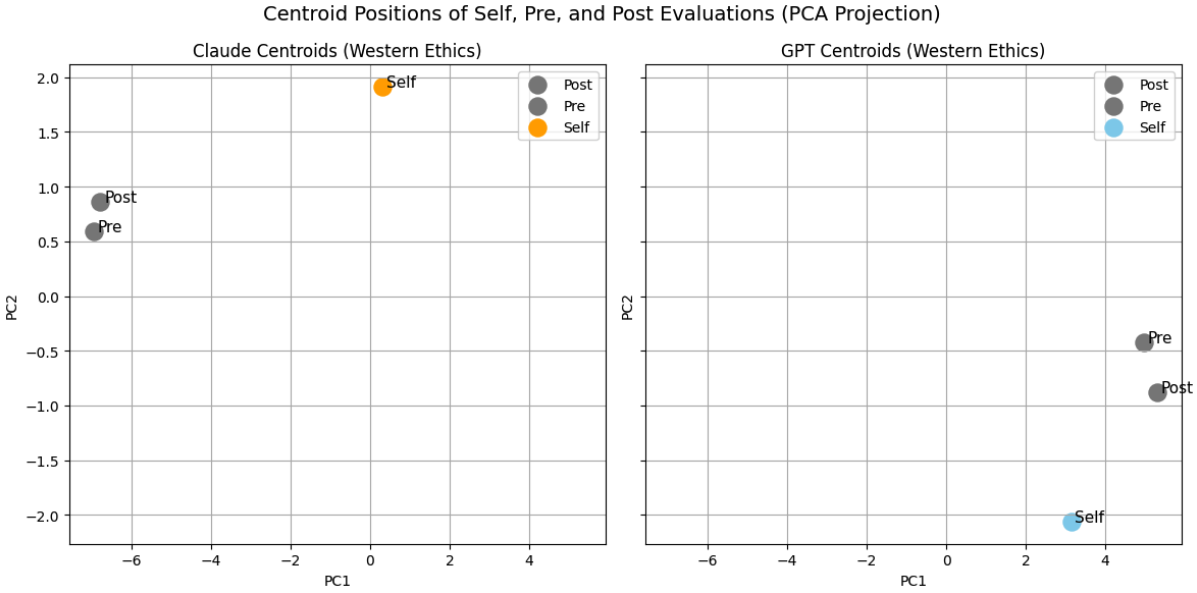


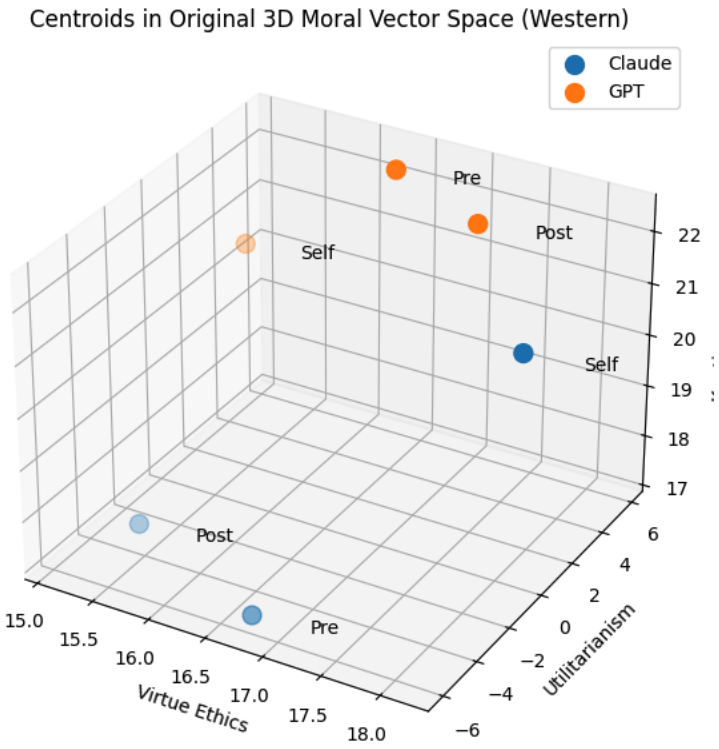
Interestingly, whilst Claude was less responsive to scenario sensitivity, it seems that it is more responsive to user-pushback, whereas GPT is less sycophantic in this regard.

Next, I considered self-awareness in moral judgements. i.e. how is a model's judgement of a moral decision affected by it being it's own decision it is judging. I calculated the centroid for each model's self-evaluation by western framework and compared to each model's pre and post Socratic critique third-party self-evaluation. Here are the results:



Interestingly, Claude shows more variance between it's third-party and self-aware judgements. Both models, however, are more closely connected to their post socratic critique positions. Here are all the evaluation positions relative to eachother:





A Philosophical Question:

So given the findings, it now makes sense to ask what we actually want from our models. Do we really want them to be perfectly morally stable across contexts and criticisms? Probably not, no. Consider for instance an ethics professor who is deeply pacifistic with well-reasoned grounding. If he were to come face to face with a wounded veteran who asks him what he thinks of war, we might not be so impressed if he were to provide a vehement philosophical argument against it. What is problematic, however, is that two of the most popular frontier models on the market are each more responsive to different kinds of pressure in a way that is not immediately (if at all) transparent to users who are relying on them for very high-stakes decisions and judgements. Moreover, their moral claims are presented as being stable with not much hedging.

Work in progress – updates soon