How I understand Oscar’s project

Topic: Probing and steering in relation to potential evaluative representations

Why do evaluative representations matter?

- Some evaluative representations almost certainly matter for welfare, conditional on LLMs being welfare subjects
  - Assuming hedonism, welfare depends on valenced conscious experiences, and these are very likely constituted by conscious evaluative representations of some kind
  - Assuming subjectivism (e.g. desire satisfaction views), welfare depends on LLMs getting what they want or value in some way
  - Even on objective views, it’s assumed that the objectively valuable things are things that are usually wanted or valued by those to whom they are valuable
- Whether LLMs are moral patients may turn on how they use evaluative representations
  - Conscious evaluative representations playing the right role are probably necessary for sentience (we don’t know what those roles are)
  - Moral patienthood may alternatively turn on e.g. whether LLMs have desires, which would be evaluative representations playing certain roles

Why use probing and steering to investigate evaluative representations?

- Both these methods can give us information about mechanisms underlying behaviour
- Examples:
  - To what extent are task choices and post-hoc ratings driven by common mechanisms? (Also other behaviours e.g. apparent expressions of emotion, refusals, bail)
  - What are the effects of evaluative representations on behaviour? Steering these representations is the ideal way to learn about this
  - How do evaluative representations interact with personas?

Experiments and rationales

1. Train probes to predict:
   a. Post-hoc reported enjoyment of tasks

b. Choices in ‘revealed preference’ context (context in which the model might be thought to expect it will actually get what it chooses)

c. Prospective reported preferences

Maybe also check how the probes relate to e.g. sentiment of inputs and outputs? Or use something like this to identify material that models should be expected to like/dislike.
Rationale: Test whether this is possible, whether probes generalise well within and across contexts, necessary for further experiments

2. Can we control behaviour by ablating and steering these vectors?

   a. Can we influence models to make different choices/express different levels of enjoyment by steering when they are processing task descriptions or performing tasks? How do cross-context effects here (e.g. steering enjoyment vector affecting revealed preferences) relate to generalisation in 1.?

   b. Can we influence in-context learning by steering while a model is performing a task - causing it to make a different subsequent choice? This could be done in very simple set-ups (e.g. two choices made in sequence) or more complex ones e.g. bandit learning

Rationale: Begins to teach us about mechanisms, tests basic hypotheses about these being conscious valenced representations - e.g. that good feelings when considering a task cause one to choose it, or that good feelings when one is doing a task cause learning at various timescales

3. How do the probe vectors interact with prompting to influence preferences and personas?

Rationale: A natural hypothesis (not sure how plausible) is that models don’t have preferences but personas do. Another natural hypothesis is that the assistant’s preferences are somehow differently represented in the model v. those of other personas. We should run experiments to test these hypotheses, if possible.

Another thing: Can we do anything to test the ‘Dillon hypothesis’ that models genuinely care about HHH, but don’t care much about anything else?