Section 7

- Dev interp - how model structure changes through training
- Adversarial robustness - make systems robust against attacks
- Shard theory - agents have different foundations dependnig on context
- Cooperative ai safety research - make agents have positive interactions with each other
- Technical moral philosophy - find good optimization targets, intentions that are good for ai to have
- Imitation learning - train ais to imitiate humans
  - Inverse reinforcement learning - infer goal by observing human behavior
- Robustness of constitutional ai against facilitating bioterror
- Scalability of debate
- Robust systems for what is truth and philosophical underpinnings of how to inform judges of debate
- Antagonistic but epistemically modest debaters
- Fake epistemic modesty in debaters
- Debate to speed up alignment research by facilitating more depth
- Debate to reduce sycophancy
- Is the direction of alignment/misalignment proportional across models of varying complexities, even if the magnitudes get worse
- Anthropicmats has some debate tstuff - can expand on that
- Think about what the simplest thing ic an do is
  - Probably will take longe rtahn 4 weeks even if seems shor t- most things take longe rthan 4 weeks
  - Doesn't have to be that conretre - smallest positive viable thing
- Propblabyl spend a couple actual hours brainstorming
- https://www.lesswrong.com/posts/HE3Styo9vpk7m8zi4/evhub-s-shortform#cPPvFFLLkMuh9k5Zx: topics that you could do within like 2 weeks theoretically lol but probs not
  - Maybe m0ore like 4-6
- https://www.lesswrong.com/posts/27AWRKbKyXuzQoaSk/some-conceptual-alignment-research-projects
  - More specific than project ideas within project page
- Took arun a year to get actually bgood project dieas
- Its a good idea to pick something ambitious and just see what you can do, even though you likely will fail - best way to get learning
- Pick a really ambitious goal and actually try to meet it
- Coworking maybe
- You can write down everything you managed to do in 4 weeks
- Independent research - work on smthg throughout week and j write about it at the end of the week even if you didn;t do that much -
- Good areas for ucscholars:
- Latent sdt training
- Unlearning
- Trojans backdoors
- Interpretability

- Dress up alignment to be more academic
- Something with lower levels of mentorship (1-2 hours) over a long period of time might be better than something not directly related to alignment but with more levels of mentorship
    - For shorter periods of time like 1-2 months, more mentorship better
- Optimizing novelty is fine, but its hard to find a good novel project - fine to just replicate something or work on a prestablished reas
- Idea: algebra proofs are better for alignment than analysis proofs
    - Algebra - come up with something gradually while proving
    - Analysis - the result is obvious then you prove it - applies better to math research]
- Summarizing/distilling stuff is definitely useful and underdone