Some conceptual alignment research projects

174
Ω 71

by **Richard_Ngo** 25th Aug 2022

Research Agendas | AI | Curated

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Some research outputs I'd love to see, focused on exploring, clarifying and formalizing important alignment concepts. I expect that most of these will be pretty time-consuming, but happy to discuss for people who want to try:

1. A paper which does for deceptive alignment what the goal misgeneralization paper does for inner alignment, i.e. describing it in ML language and setting up toy examples (for example, telling GPT-3 to take actions which minimize changes in its weights, given that it's being trained using actor-critic RL with a certain advantage function, and seeing if it knows how to do so).
2. A paper which does the same for gradient hacking, e.g. taking these examples° and putting them into more formal ML language.
3. A list of papers that are particularly useful for new research engineers to replicate.
4. A takeover scenario which covers all the key points in https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/, but not phrased as an argument, just phrased as a possible scenario (I think you can't really make the argument rigorously in that little space).
5. A paper which defines the concepts of implicit planning, implicit value functions, implicit reward models, etc, in ML terms. Kinda like https://arxiv.org/abs/1901.03559 but more AGI-focused. I want to be able to ask people "does GPT-3 choose actions using an implicit value function?" and then be able to point them to this paper to rigorously define what I mean. I discuss this briefly in the phase 1 section here°.
6. A blog post which describes in as much detail as possible what our current "throw the kitchen sink at it" alignment strategy would look like. (I'll probably put my version of this online soon but would love others too).
7. A blog post explaining "debate on weights°" more thoroughly.

8. A blog post exploring how fast we should expect a forward pass to be for the first AGIs - e.g. will it actually be slower than human thinking, as discussed in this comment°?
9. A blog post exploring considerations for why model goals may or may not be much more robust to SGD than model beliefs, as discussed in framing 3 here°. (See also this paper on gradient starvation - h/t Quintin Pope; and the concept of persistence to gradient descent discussed here.)
10. A blog post explaining why the "uncertainty" part of CIRL only does useful work insofar as we have an accurate model of the human policy, and why this is basically just as hard as having an accurate model of human preferences.
11. A blog post explaining what practical implications Stuart Armstrong's impossibility result has.
12. As many alignment exercises as possible to help people learn to think about this stuff (mine aren't great° but I haven't seen better).
13. A paper properly formulating instrumental convergence, generalization to large-scale goals, etc, as inductive biases in the ML sense (I do this briefly in phase 3 here°).
14. A mathematical comparison between off-policy RL and imitation learning, exploring ways in which they're similar and different, and possible algorithms in between.
15. A blog post explaining the core argument for why detecting adversarially-generated inputs is likely much easier than generating them, and arguments for why adversarial training might nevertheless be valuable for alignment.
16. A blog post exploring the incentives which models might have when they're simultaneously trained to make predictions and to take actions in an RL setting (e.g. models trained using RL via sequence modeling).
17. A blog post exploring pros and cons of making misalignment datasets for use as a metric of alignment (alignment = how much training on the misalignment dataset is needed to make it misaligned).
18. A paper providing an RL formalism in which reward functions can depend on weights and/or activations directly, and demonstrating a simple but non-trivial example.
19. A blog post evaluating reasons to think that situational awareness° will be a gradual development in models, versus a sharp transition.

20. A blog post explaining reasons to expect capabilities to be correlated with alignment while models lack situational awareness, and then less correlated afterwards, rather than the correlation continuing.
21. A blog post estimating how many bits of optimization towards real-world goals could arise from various aspects of a supervised training program (especially ones which slightly break the cartesian formalisms) - e.g. hyperparameter tuning, many random seeds, training on data generated by other AIs, etc.
22. A sketch of what a model-free version of AIXI would look like (according to one person I talked to, it's a lot like decision transformers).
23. A blog post evaluating whether shard theory° makes sense/makes novel predictions compared with Steve Byrnes' model of the brain° (he partly explains this in a comment on the post, but I'm still a bit confused).
24. A blog post or paper reviewing what types of feedback humans perform best and worst at (e.g. reward vs value feedback) and then designing a realistic setup for optimal-quality human feedback.
25. A blog post compiling examples of surprising emergent capabilities (especially in large language models).
26. An investigation of the extent to which human concept representations are localized to individual neurons, versus being spread out across different neurons.

Research Agendas 4 | AI 4 | Curated

Mentioned in

76 Resources that (I think) new alignment researchers should know about
57 EA & LW Forums Weekly Summary (21 Aug - 27 Aug 22')
46 Levelling Up in AI Safety Research Engineering
40 List of technical AI safety exercises and projects
10 Request for Alignment Research Project Recommendations

Load More (5/6)

16 comments, sorted by top scoring

[-] **geoffreymiller** 1y 14 0

Regarding #23, I'm working on a friendly critique of shard theory, but it won't be ready to share for a few weeks.

Preview: as currently framed, shard theory seems to involve a fairly fundamental misconception about the nature of genotype-phenotype mappings and the way that brain systems evolve, with the result that it radically underestimates the diversity, complexity, and adaptiveness of our evolved motivations, preferences, and values.

In other words, it prematurely rejects the 'massive modularity' thesis of evolutionary psychology, and it largely ignores the last three decades of research on the adaptive design details of human emotions and motivations.

I think it'll be important for AI alignment researchers (and AI systems themselves) to take evolutionary biology and evolutionary psychology more seriously in trying to understand and model human nature and human preferences. (But then, I'm possibly biased, since I've been doing machine learning research since the late 1980s, and evolutionary psychology research since the early 90s....)

[–] **cousin_it** 1y Ω 4 13 9

> A takeover scenario which covers all the key points in https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/, but not phrased as an argument, just phrased as a possible scenario

For what it's worth, I don't think AI takeover will look like war.

The first order of business for any AI waking up won't be dealing with us; it will be dealing with other possible AIs that might've woken up slightly earlier or later. This needs to be done very fast and it's ok to take some risk doing it. Basically, covert takeover of the internet in the first hours.

After that, it seems easiest to exploit humanity for awhile instead of fighting it. People are pretty manipulable. Here's a thought: present to them a picture of a thriving upload society, and manipulate social media to make people agree that these uploads smiling on screens are really conscious and thriving. (Which they aren't, of course.) If done right, this can convince most of humanity to make things as nice as possible for the upload society (ie build more computers for the AI) and then upload themselves (ie die). In the meanwhile the "uploads" (actually the AI) take most human jobs, seamlessly assuming control of civilization and all its capabilities. Human stragglers who don't buy the story can be called anti-upload bigots, deprived of tech, pushed out of sight by media control, and eventually killed off.

[–] **jacob_cannell** 1y 15 3

Yes, this is very similar to my main worst case scenario. It is interesting/terrifying specifically because it is intentionally similar to the best case scenario.

[–] **phone.spinning** 1y 4 0

I think AI of the capability level that you describe will either already have little need to exploit people, or will quickly train successors that wouldn't benefit from this. I do think deception is a big issue, but I think the important parts of deception will be earlier in terms of AI capability than you describe.

[–] **Raemon** 1y Ω 2 5 1

Curated. I think shovel-ready projects that can help with alignment are quite helpful for the field, in particular right now when we have a bunch of smart people showing up, looking to contribute.

[–] **David Johnston** 1y 4 1

This is an awesome list, thanks.

[–] **watermark** 3mo 3 0

Another potential idea:

27. A paper which does for the sharp left turn° what the goal misgeneralization paper does for inner alignment (or, at least, breaking the SLT into sub-problems and making a paper for one of the sub-problems)

[–] **Ramana Kumar** 1y Ω 2 3 0

This post° (and the comment° it links to) does some of the work of #10. I agree there's more to be said directly though.

[–] **307th** 1y 2 3

1. A paper which does for deceptive alignment what the goal misgeneralization paper does for inner alignment, i.e. describing it in ML language and setting up toy examples (for example, telling GPT-3 to take actions which minimize changes in its weights, given that it's being trained using actor-critic RL with a certain advantage function, and seeing if it knows how to do so).

If I'm understanding this one right OpenAI did something similar to this for purely pragmatic reasons with VPT, a minecraft agent. They first trained a "foundation model" to imitate humans playing. Then, since their chosen task was to craft a diamond pickaxe, they finetuned it with RL, giving it reward for each step along the path to a diamond pickaxe that it successfully took. There's a problem with this approach:

"A major problem when fine-tuning with RL is catastrophic forgetting because previously learned skills can be lost before their value is realized. For instance, while our VPT foundation model never exhibits the entire sequence of behaviors required to smelt iron zero-shot, it *did* train on examples of players smelting with furnaces. It therefore may have some latent ability to smelt iron once the many prerequisites to do so have been performed. To combat the catastrophic forgetting of latent skills such that they can continually improve exploration throughout RL fine-tuning, we add an auxiliary Kullback-Leibler (KL) divergence loss between the RL model and the frozen pretrained policy."

In other words they first trained it to imitate humans, and saved this model; after that they trained the agent to maximize reward, but used the saved model and penalized it for deviating too far from what that human-imitator model would do. Basically saying "try to maximize reward but don't get too weird about it."

[–] **ukc10014** 10mo 1 0

I've taken a crack at #4 but it is more about thinking through how 'hundreds of millions of AIs' might be deployed in a world that looks, economically and geopolitically, something like today's (i.e. the argument in the OP is for 2036 so this seems a reasonable thing to do). It is presented as a flowchart° which is more succinct than my earlier longish post°.

[–] **Martin Vlach** 1y 1 0

"goal misgeneralization paper does does"–typo

"list of papers that it's"→ that are

[–] **Richard_Ngo** 1y 3 1

Thanks! Fixed, although I feel confused about whether the second is actually grammatically incorrect or not (you could interpret it as "papers such that it's good to replicate those papers", but idk what the rulebook would say).

[–] **Joe_Collman** 1y Ω 1 1 0

Thanks for the list.

On (7), I'm not clear how this is useful unless we assume that debaters aren't deceptively aligned. (if debaters are deceptively aligned, they lose the incentive to expose internal weaknesses in their opponent, so the capability to do so doesn't achieve much)

In principle, debate on weights could add something over interpretability tools alone, since optimal in-the-spirit-of-things debate play would involve *building* any useful interpretability tools which didn't already exist. (optimal not-in-the-spirit-of-things play likely involves persuasion/manipulation)

However, assuming debaters are not deceptively aligned seems to assume away the hard part of the problem. Do you expect that there's a way to make use of this approach before deceptive alignment shows up, or is there something I'm missing?

[–] **Richard_Ngo** 1y Ω 16 35 25

There's one attitude towards alignment techniques which is something like "do they prevent all catastrophic misalignment?" And there's another which is more like "do they push out the frontier of how advanced our agents need to be before we get catastrophic misalignment?" I don't think the former approach is very productive, because by that standard no work is ever useful. So I tend to focus on the latter, with the theory of victory being "push the frontier far enough that we get a virtuous cycle of automating alignment work".

[–] **Joe_Collman** 1y Ω 2 2 0

Ok, thanks - I can at least see where you're coming from then.
Do you think debate satisfies the latter directly - or is the frontier only pushed out if it helps in the automation process? Presumably the latter (??) - or do you expect e.g. catastrophic out-with-a-whimper dynamics before deceptive alignment?

I suppose I'm usually thinking of a "do we learn anything about what a scalable alignment approach would look like?" framing. Debate doesn't seem to get us much there (whether for scalable ELK, or scalable anything else), unless we can do the automating alignment work thing - and I'm quite sceptical of that (weakly, and with hand-waving).

I can buy a safe-AI-automates-quite-a-bit-of-busy-work argument; once we're talking about AI that's itself coming up with significant alignment breakthroughs, I have my doubts. What we want seems to be unhelpfully

complex, such that I expect we need to target our helper AIs precisely (automating *capabilities* research seems much simpler).

Since we currently can't target our AIs precisely, I imagine we'd use (amplified-)human-approval. My worry is that errors/noise compounds for indirect feedback (e.g. deep debate tree on a non-crisp question), and that direct feedback is only as good as our (non-amplified) ability to do alignment research.
AI that doesn't have these problems seems to be the already dangerous variety (e.g. AI that can infer our goal and optimize for it).

I'd be more optimistic if I thought we were at/close-to a stage where we know the crisp high-level problems we need to solve, and could ask AI assistants for solutions to those crisp problems.

That said, I certainly think it's worth thinking about how we might get to a "virtuous cycle of automating alignment work". It just seems to me that it's bottlenecked on the same thing as our direct attempts to tackle the problem: our depth of understanding.

[+] [comment deleted] 1y 2 0

Deleted by joshc, 08/29/2022
Reason: misinterpreted Ngo's point

Moderation Log