How promising is automating alignment research? (literature review)

Summary

Automating alignment research is one approach to alignment that has gained much more visibility with the Open AI superalignment plan announcement. Some of automating alignment's selling points (if successful) include potentially resulting in an enormous amount of alignment research even in short calendar time and the apparent relative ease of (even if non-robustly) aligning systems similar to the current state-of-the-art (e.g. large language models, foundation models; see this presentation of mine for many more details), which could be used as automated alignment researchers/research assistants. Notably, if successful, automating alignment research could plausibly be the most scalable alignment research agenda (and probably by a wide margin). At the same time, strategies to automate alignment research like the superalignment plan have received a lot of criticism within the alignment community (see e.g. this post).

This project aims to get more grounding into how promising automating alignment research is as a strategy, with respect to both advantages and potential pitfalls, with the superalignment plan as a potential blueprint/example (though ideally the findings would apply more broadly). This will be achieved by reviewing, distilling and integrating relevant research from multiple areas/domains[1], with a particular focus on the science of deep learning and on empirical findings in deep learning and language modelling (see my presentation for examples of what this might look like/for a potential starting point). Depending on team members' profiles, this could expand much more broadly, covering e.g. reviewing and distilling relevant literature from AI governance, multidisciplinary intersections (e.g. neuroscience and alignment), relevant predictions on prediction markets, and the promise of automating larger parts of AI risk mitigation research (e.g. including AI governance research).

This could also inform e.g. how promising it might be to start more automated alignment/AI risk mitigation projects or to dedicate more resources to existing ones.

The non-summary

This section is written in a Q&A format.

---

[1]The closer in the future automating alignment research becomes possible (e.g. the superalignment plan's deadline is in less than 4 years), the more likely it is that the systems used look like those of today, so the more likely it is that theoretical and empirical results about them would still apply. Additionally, some theoretical results are more likely to hold more generally, e.g. potentially this parallelism - expressivity tradeoff conjectured in The Parallelism Tradeoff: Limitations of Log-Precision Transformers.

- **If the project succeeds, how would this be useful for reducing risk from AGI/TAI?**

  I will discuss how this project would be useful for reducing *existential* risk (x-risk) from AGI/TAI and break this into 2 parts:
  - How much is x-risk reduced if automated alignment succeeds?
  - How much can this specific project increase the likelihood of automating alignment succeeding / reduce AGI/TAI x-risk more broadly?

  How much is x-risk reduced if automated alignment succeeds?

  As of November 10th, 2023 this Metaculus prediction estimates that if OpenAI announced 'that it has solved the core technical challenges of superintelligence alignment by June 30, 2027', then the probability of humans going extinct by 2100 would decrease from 10% to 3%. This suggests that, if this prediction were taken at face value, then automating alignment research is probably easily one of the most impactful projects to work on for reducing x-risk.

  On a more personal note, automated alignment seems to me like probably one of the most efficient and scalable ways to significantly reduce AGI/TAI risk. Its scalability seems better than that of any other alignment research agenda, since if successful it seems likely to lead to the equivalent of millions of person-years of alignment research which could be performed within a couple of months; corresponding to more automated alignment research than all the previous alignment research done by non-augmented humans. Thus, in principle, building an aligned human-level automated alignment alignment researcher allows for fully delegating the problem of aligning superintelligence. Related, automating alignment research seems especially promising in worlds in which aligning ~human-level systems is relatively easy and we can get at least a few months between time with ~human-level fully-automated alignment researcher and time of significant probability of existential catastrophe. The current progress in AI governance (including e.g. on evals and measures against misuse coming out from the recent AI safety summit) seems to me to favour the promise of this approach (though ideally we would prefer for much more coordination and caution and less AI race conditions). Automating alignment research using systems similar to the current ones should also have the comparative advantage (vs. other alignment approaches) that many actors should be incentivized to align current systems to be helpful even for just purely commercial reasons; this should significantly help with trying to build aligned automated alignment researchers that look like current systems (e.g. to plug into the superalignment plan). These favorable incentives don't apply as clearly to plans which try to directly align systems which might appear later and whose shape is less clear (e.g. superintelligence).

  Please see my slides for many more details, including (more) evidence/references for some of the claims above (mostly focused on reasons for optimism about automated alignment approaches, with the superalignment plan as an example).

How much can this specific project increase the likelihood of automating alignment succeeding / reduce AGI/TAI x-risk more broadly?

Related, see these Metaculus and Manifold predictions about the likelihood of the superalignment plan succeeding. This project could have a direct impact on these estimates, e.g. by informing the AI safety community's beliefs and perhaps (ambitiously) even by directly influencing the likelihood of automated alignment plans succeeding.

As far as I can tell, trying to review the existing evidence about how feasible and how safe automating alignment research is (e.g. using systems similar to state-of-the-art i.e. large language models) has been quite neglected, especially in terms of integrating potential evidence from multiple fields. I also expect that at least some parts of this proposal are highly tractable compared to many other research projects/agendas; e.g., to give some intuition, reviewing / distilling existing work can often be much easier than performing the same work from scratch. More broadly, my impression has been that reviewing/distilling research from other fields has been historically neglected in the AI alignment/AI existential safety community; this project could also reduce AGI/TAI x-risk by informing other agendas (e.g. the translucent thoughts hypotheses, projects at the intersection of neuroscience and alignment).

I'd also argue that automating alignment has been comparatively neglected overall, given that OpenAI's superalignment is the only well-resourced public plan that I'm aware of. This project could increase the likelihood of more automated alignment plans existing / of more resources dedicated to automated alignment plans. For example, the UK's AI Safety Institute funding of 100M pounds / year suggests there could be more willingness from governments to spend on AI existential safety research than has been the case historically. In this context, automating AI alignment (and AI safety more broadly) could be both incredibly scalable (vs. e.g. onboarding new [human] scientists) and very cost-effective (e.g. see GPT-4's API costs). Positive findings from this project could inform AI safety funders (including historical AI safety funders like Open Philanthropy) and for example increase their confidence that automating alignment research could pass their funding bars and be both uniquely scalable and incredibly cost-effective. On the other hand, negative findings could also be very important e.g. with respect to how to prioritize alignment approaches and updates on where alignment people should work/which plans to support, etc.

- **What are the steps you need to complete to finish this project?**

  The steps might look something like (this will also depend on team members' profiles, e.g. how much generalist vs. specialist, what areas of specialization):

  - Brainstorm [more] potentially relevant literatures (domains of knowledge); I've already done at least some of this, though not necessarily always very systematically.
  - Literature search (e.g. citation tracking), especially reviews of such literatures; I've already done a significant portion of this.
  - Summarize the relevant literatures.

    - Bring together the acquired insights in writeups.

  These steps could be iterated multiple times (loop the above steps); as more insights are found, the focus of the investigation itself could change. Finally, we would work on the final writeup.

- **What's the first step?**

  The first step might look like brainstorming [more] potentially relevant domains for scientific literatures to review or like brainstorming of potential crucial considerations. This will also depend on the team members' profiles.

- **What can go wrong, and what's the backup plan if that happens?**

  As in most projects, many things can go wrong, from planning fallacy, to coordination challenges, to sunk cost fallacy, etc. Perhaps one of the more significant and less generic failure modes (skipping risks and downsides, covered in another section) would be missing some very crucial consideration, or only realizing it too late in the project, when we could have thought of it much sooner. The impact of this project could also become counterfactually much weaker if e.g. large language models become very good at this kind of distillation-focused research very soon.

- **What's the *most* ambitious version of this project?**

  The most ambitious version of this project probably looks something like doing very comprehensive reviews of multiple [sub]domains relevant to automating alignment (and potentially AI safety more broadly), spanning e.g. technical AI alignment, AI governance, theoretical ML, empirical ML results, relevant forecasts [on prediction markets], etc. This would allow us to come up with crucial considerations relevant to the feasibility and safety of automating alignment research (and maybe, along the way, crucial considerations about other alignment agendas e.g. the translucent thoughts hypothesis). Another potential expansion of the project is to evaluate the promise of automated all AI safety/risk research. This could potentially significantly inform the superalignment plan or potentially lead to more (and perhaps broader) similar automated alignment research plans.

- **What's the *least* ambitious version of this project?**

  The least ambitious version looks something like only focusing on one narrow subdomain and a relatively small number of papers which seem like they might be particularly relevant to automated alignment - e.g. recent theoretical ML results relevant to ML transparency and risks of deception.

- **What's your scope? What research do you *not* focus on?**

  Running experiments is out of scope. Also out of scope will be, most likely, areas of Machine Learning conceptually far from / less competitive with the foundation models paradigm.

Output

Part of the format of AISC is that projects have a beginning and an end. At the end of the project, what will you have produced?

A blogpost? An academic paper? A github repo? Something else?

At the end of the project, I expect we will have produced one (or multiple, e.g. a sequence of) blog post (on LessWrong / the Alignment Forum) and potentially an academic preprint/paper (e.g. on arXiv).

Risks and downsides

Does your project have any risk or other potential downsides? E.g. infohazards, potential AI capabilities progress, etc.

This project might have some risks of infohazards by e.g. also reviewing capabilities-relevant ML literature. These seem relatively low since we'll focus on how the ML literature is relevant to automated alignment research and we expect to mainly release this in alignment-relevant fora (LessWrong/the Alignment Forum), but we'll pay extra attention to infohazard risks too.

Acknowledgements

This project has been significantly influenced by the work of many researchers working on automating alignment research and related topics, such as (among others) Jan Leike, Ajeya Cotra and Lukas Finnveden.
I am grateful for comments, feedback and discussions on this topic to the participants of AI safety bootcamps (ML4G France 3, ML4G Germany, ML4G Switzerland 2) and at talks I gave on earlier work that influenced this proposal (at Oxford Trajan House and at Center on Long-Term Risk) and to my flatmates from Duality House, London. Special thanks to Charbel-Raphael Segerie, who has given extensive feedback and with whom I've had many helpful conversations on this topic, and to Linda Linsefors, for advice and help with drafting this proposal for AISC.
Most of the work which subsequently turned into this proposal was carried out while I was funded by Center on Long-Term Risk (CLR).

Team

Team size

As a result of its scope and potential multidisciplinarity (also see the 'Skill requirements' section below), I expect this project can accommodate > 100 hrs/week, so ideally I'd prefer

to work with at least 4 additional people with at least 20 hrs/week availability, but I could see anywhere up to e.g. 300 hrs/week in total being useful.
All else equal, I would prefer to work with team members with longer hrs/week availability.

Research Lead

Bogdan-Ionut Cirstea
cirstea.bogdanionut@gmail.com

I have very significant experience in ML (PhD and postdoc) and AI alignment research (~1 year full time, a couple of additional years part-time) and in AI alignment field-building (~ 2.5 years part-time). Doing (conceptual) alignment facilitating for AGISF and for AI safety bootcamps has also allowed me to have a very broad view of alignment research, which I expect to come in very handy for this project. Please see my CV for more details.
I am also strongly motivated by this project and expect it to be my main focus for the near-term at the very least.

I expect to easily spend at least 10 hrs/week (most likely > 20 hrs/week) and certainly at least 3 hrs/week even in “worst case scenarios”, e.g. I get hired to work on something else. As an independent researcher, this is currently my main research project and will be the largest part of my next round of applying for independent funding.

Team Coordinator

The TC is the ops person of the team. They are in charge of making sure meetings are scheduled, check in with individuals on their task progress, etc. The job of the TC is important, but not expected to take much time (except for project management-heavy teams). TC and RL can be the same person.

I’d prefer for someone else (who would ideally be highly involved in this project, at least 20 hrs/week) to take on the TC role. I can also provide some help with this, especially at the high (strategic) level.

Skill requirements

This subsection is written in a Q&A format.

What skills are needed for this project?

A wide range of profiles and skills could contribute significantly to this project. Some (non-comprehensive) examples of relevant crucial considerations and relevant domain areas/example research (see slides for more details/examples on what such research might look like more concretely):

- How much time between automated alignment researchers and significant x-risk from AI misuse? Compute governance, evals and governance

- When should we expect automated alignment research vs. automated capabilities research? E.g. scaling laws (including for transfer learning), t-AGI framework, scaling laws and temporal horizons, science of DL
- How hard does aligning ~human-level foundation systems seem to be? Science of DL, empirical DL findings
    - How human-like are current systems and how hard would it be to make them more human-like for alignment purposes? See this review on comparing artificial and biological neural networks and more linkposts on my LessWrong profile

- **What minimum skills or understandings does any team member need to be able to contribute to this project?**

  Minimum ML understanding, minimum AI safety tech knowledge equivalent to having gone through AGISF, good communication (distillation) skills, basic research skills. A wide variety of additional skills could be useful, especially good distillation (strong writing skills), strong generalist skills, more advanced ML/theoretical CS/math skills.

- **What diverse skills or backgrounds would you value having in your team, even if they are hard to find? Dream big: If you could get any person with any skills, what skills would they have?**

  The ideal candidate would possess all the above-mentioned skills, as well as very strong motivation and strong time commitment (e.g. FTE).

- **Are there any skills that are needed for this project that you don't have yourself, and therefore need someone else to bring to the project?**

  More advanced theoretical computer science / math skills seem to me like the most complimentary skills to my own. I'd also strongly prefer for somebody else to handle [most] project management tasks (especially for larger team size / fewer committed work hrs per team member).