Request for proposals for projects in AI alignment that work with deep learning systems

As part of our work on reducing potential risks from advanced artificial intelligence, we are seeking proposals for projects working with deep learning systems that could help us understand and make progress on AI alignment: the problem of creating AI systems more capable than their designers that robustly try to do what their designers intended. We are interested in proposals that fit within certain research directions, described below, that we think could contribute to reducing the risks we are most concerned about.

Anyone is eligible to apply, including those working in academia, industry, or independently. Applicants are invited to submit proposals for up to $750K in total funding covering up to 2 years. Grants will cover individual projects and will not be renewed, though we may invite grantees who do outstanding work to apply for larger and longer grants in the future.

Proposals are due December 1, 2021.

Submit a proposal here[LINK].

If you have any questions, please contact ai-alignment-rfp@openphilanthropy.org.

Our view of alignment risks from advanced artificial intelligence

This section was written by Nick Beckstead and Asya Bergal, and may not be representative of the views of Open Philanthropy as a whole.

We think the research directions below would be pursued more fruitfully by researchers who understand our background views about alignment risks from advanced AI systems, and who understand why we think these research directions could help mitigate these risks.

In brief:

- We believe it is plausible that later this century, advanced AI systems will do the vast majority of productive labor more cheaply than human workers can.[1]

[1] See Cotra 2020, Davidson 2021a, Davidson 2021b, and, more broadly, Holden Karnofsky's in-progress "Most Important Century" series.

- We are worried about scenarios where AI systems more capable than humans acquire undesirable objectives that make them pursue and maintain power in unintended ways, causing humans to lose most or all influence over the future.
- We think it may be technically challenging to create powerful systems that we are highly certain have desirable objectives. If it is significantly cheaper, faster, or otherwise easier to create powerful systems that may have undesirable objectives, there may be economic and military incentives to deploy those systems instead.[2]
- We are interested in research directions that make it comparatively easier to create powerful systems that we are highly certain have desirable objectives.

In this request for proposals, we are focused on scenarios where advanced AI systems are built out of large neural networks. One approach to ensuring large neural networks have desirable objectives might be to provide them with reward signals generated by human evaluators.[3] However, such a setup could fail in multiple ways:

- **Inadequate human feedback:** It's possible that in order to train advanced AI systems with desirable objectives, we will need to provide reward signals for highly complex behaviors that have consequences that are too difficult or time-consuming for humans to evaluate.[4]
    - **Deceiving human evaluators:** It may be particularly difficult to provide good reward signals to an AI system that learns undesirable objectives during training and has a sophisticated model of humans and the training setup. Such a system may "deceive" the humans, i.e. deliberately behave in ways that appear superficially good but have undesirable consequences.
- **Competent misgeneralization:** Even if an AI system has an abundant supply of good reward signals and behaves consistently with desirable objectives on the training distribution, there could be contexts outside of the training distribution where the system retains its capabilities but pursues an undesirable objective.
    - **Deceptive misgeneralization:** Rather than subtly misbehaving during training as in "deceiving human evaluators", an sophisticated AI system that learns undesirable objectives may choose to behave in only desirable ways during training, maximizing its chances of being deployed in the real world, where it can more effectively pursue its true objectives. This case and the analogous one

[2] For this and the above bullet point, see this recent draft report by Joseph Carlsmith, an Open Philanthropy Senior Research Analyst, which looks at these scenarios in more detail. To get additional perspectives on possible scenarios, it may also be useful to read:

- Posts by Paul Christiano here and here.
- *Superintelligence* by Nick Bostrom,
- *Human Compatible* by Stuart Russell,

[3] There may be viable approaches to this problem that do not rely on human feedback; they are not the focus of this request for proposals.

[4] Alternatively, we could hope to be able to use features of the network's internals or training setup to argue that its objectives won't become undesirable, even as humans aren't able to provide additional reward.

above may pose special challenges because of the adversarial relationship between the system and its designers.[5]

Research directions

We are soliciting proposals that fit within one of the following research directions. For each research direction, we give a brief description below and link to a document describing the direction in depth.

Direction 1: Measuring and forecasting risks

Proposals that fit within this direction should aim to measure concrete risks related to the failures we are worried about, such as reward hacking,[6] misgeneralized policies, and unexpected emergent capabilities. We are especially interested in understanding the trajectory of risks as systems continue to improve, as well as any risks that might suddenly manifest on a global scale with limited time to react. We think this research direction could allow us to better direct future research, as well as to make stronger arguments for worrying about certain risks.

Direction 2: Techniques for enhancing human feedback

Proposals that fit within this direction should aim to address the inadequate feedback problem by developing general techniques for generating good reward signals using human feedback that could apply to settings where it would otherwise be prohibitively difficult, expensive, or time-consuming to provide good reward signals. We are especially interested in proposals that use these techniques to train models to complete tasks that would otherwise be difficult to accomplish.

Direction 3: Interpretability[LINK]

---

[5] Nick Bostrom describes this failure mode in *Superintelligence*, p. 117:
"…one idea for how to ensure superintelligence safety… is that we validate the safety of a superintelligent AI empirically by observing its behavior while it is in a controlled, limited environment (a "sandbox") and that we only let the AI out of the box if we see it behaving in a friendly, cooperative, responsible manner. The flaw in this idea is that behaving nicely while in the box is a convergent instrumental goal for friendly and unfriendly AIs alike. An unfriendly AI of sufficient intelligence realizes that its unfriendly final goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual."

Additional discussions of the possibility of such failure modes can be found in Hubinger et al.'s Risks from Learned Optimization in in Advanced Machine Learning Systems (section 4, "Deceptive Alignment") and Luke Muelhauser's post, "Treacherous turns in the wild".

[6] "Reward hacking" refers to AI systems finding decisions that do well according to the explicit reward function, but that were unintended and undesired-- an instance of the "inadequate feedback" failure described above.

Proposals that fit within this direction should aim to contribute to the mechanistic understanding of neural networks, which could help us discover unanticipated failure modes and ensure that large models in the future won't pursue undesirable objectives in contexts not included in the training distribution (cf. "competent misgeneralization" above). Potential projects in this direction could consist of mapping small-scale structures in neural networks to human understandable algorithms, finding large-scale structures that simplify the understanding of neural networks, and learning about neurons that respond to multiple unrelated features, among others. Proposals related to scaling mechanistic interpretability to larger models are of particular interest.

Direction 4: Truthful and honest AI[LINK]
Proposals that fit within this direction should aim to contribute to the development of AI systems that have good performance on standard benchmarks while being "truthful", i.e. avoiding saying things that are false, and "honest", i.e. accurately reporting what they believe.

TODO: Add other reasons why this is good
Making models truthful and honest while achieving good performance on standard benchmarks could also teach us something about the broader problem of making AI systems that avoid certain kinds of failures while staying competitive and performant.

Potential projects could aim to develop definitions and concepts that are fruitful for relevant ML research, create benchmarks or tasks to measure truthfulness or honesty, or develop techniques for making systems that are more truthful and honest.

(in progress)

Application process

Use **this form[LINK]** to submit a project proposal.

The form asks for:

- An up-to-date CV
- A less than 10 page project description, which should include:
    - a) An outline of the proposed steps for your project, to the best of your ability, including any experiments you want to run, though we expect that many details will be uncertain until the project is underway.
    - b) A description of the outcome you are hoping for: what would we learn or gain from this project if it went well?
    - c) An explanation of impact: how would the outcome given in b) help us avoid the inadequate feedback or misgeneralization failures described above, or otherwise reduce the chance that power-seeking AI systems cause humanity to lose most or all influence over the future?

We think applicants should spend most of the proposal answering (a) and (b); however, it's important to us that the answers to (c) make sense and we will examine them critically.
- An estimated budget
- An estimated project duration

By default, we expect proposals to request no more than $750K total and to cover projects lasting no more than 2 years. If you are submitting a larger proposal, please include an explanation in your project description of why your work cannot be scoped into this budget and timeframe. Grants will cover individual projects and will not be renewed, though we may invite grantees who do outstanding work to apply for larger and longer grants in the future.

All grantees are required to submit a 3-page progress report to us every 6 months after their grant is awarded, and a final report to us after the project is finished.

Proposals are due **December 1.**

We plan to evaluate proposals in two stages. We will let applicants know if they have passed Stage 1 by early February. If you pass Stage 1, we may contact you with additional follow-up questions or ask you to join us for an interview. We anticipate making final decisions by early March.

Draft application form

Email *
Valid email address

First name *
Short answer

Last name *
Short answer

Resume or CV (PDF) *
File upload

[Optional] Institution
Short answer

[Optional] Other collaborators
Short answer

Research direction *

Multiple choice

- *Measuring and forecasting risks*
- *Techniques for enhancing human feedback*
- *<others>*

Project description (max. 10 pages)*

Please attach a document no longer than 10 pages with a description of your proposed project. Include a) An outline of the proposed steps for your project, to the best of your ability, including any experiments you want to run, though we expect that many details will be uncertain until the project is underway; b) A description of the outcome you are hoping for: what would we learn or gain from this project if it went well?; and c) An explanation of impact: how would the outcome given in b) help us avoid the inadequate feedback or misgeneralization failures described above, or otherwise reduce the chance that power-seeking AI systems cause humanity to lose most or all influence over the future? We think applicants should spend most of the proposal answering (a) and (b); however, it's important to us that answers to (c) make sense, and we will examine them critically.

Budget *

Please attach an estimated budget for your project. Be sure to include individual salaries, the cost of any computational resources required, the cost of any data collection you would want to pay for, and any indirect costs ('overhead'), if applicable. For universities, indirect costs may not exceed 10% of total direct costs. We are open to and encourage applications that make larger use of computational resources than typical academic grant applications.
File upload

Duration *

How long do you expect this project will take? It's okay if you are very uncertain-- please give a range as well as a central estimate.
Short answer

[Optional] References

Please give the names and contact information for one or more references that could speak to your ability to execute the project above. We especially value references from others who have done work in AI alignment.
Short answer

[Optional] Is there anything else you would like to share with us?

Long answer