Funding for alignment research

Paul Christiano, August 2017

(**ETA**: I am no longer making grants.)

I plan to allocate about $120k of funding for independent AI alignment research this year.

I expect funds to cover full-time work for 1-12 months at $10k/month. For particularly compelling applications it might cover longer time periods or a higher monthly rate.

I am also open to funding part-time work, or for other schemes for using money to advance our understanding of AI alignment.

Recipients will probably be paid as a contractor by a non-profit you haven't heard of. This funding is not at all associated with or endorsed by any organizations that happen to employ me as a researcher or advisor.

Application and decision-making process

Initial applications should be less than a page and can be as short as a paragraph, explaining (a) what research you would like to do, or directions you would like to consider, (b) any evidence that you can do high-quality research, especially pointers to existing papers, implementations, blog posts, thoughtful comments, or whatever else, (c) your timeline, desired length/amount of funding, current employment situation, and anything else relevant to your interest.

Email applications to funding@ai-alignment.com.

I will likely make another block of decisions in August (I allocated a first round of funding last October and another in April, and have so far allocated $70k). Depending on the applications received, there may be a follow-up discussion before a final decision.

The money will be spent on whatever projects I believe will be most helpful for alignment. My process will not be accountable to anyone else, and I will not be making an effort to be exhaustive. The funded projects are likely to be in areas that I already understand, since those are easiest for me to evaluate.

I think many people shouldn't take this kind of funding and should instead pursue more traditional career opportunities. But that's hard to evaluate from the outside, and I'm not going to be paternalistic about it.

This process is subject to change without notice up until the point when people actually receive funding.

Topics

By “AI alignment” I mean building AI systems which robustly advance human interests. More specifically, this funding is targeted at:

- *Differential progress*, which advances AI alignment *more* than it advances AI in general.
- *Existential risk* caused by AI having irreversible unintended effects on society’s trajectory (rather than causing short-term problems, or interacting with other technologies that pose an existential risk)

I expect this funding to either go to topics which fall outside of typical research paradigms in machine learning and AI, or to researchers with unusual backgrounds. That said, I won’t dismiss a project because it is too traditional.

There will naturally be a bias towards topics that I consider important. Here are a few directions that seem interesting to me (and particularly likely to be neglected), though this list is by no means exhaustive:

- Informed oversight: how can we train an AI to perform actions, and to provide information that will help an overseer evaluate it?
- Amplification: can we define a process which uses a number of “pretty good, pretty smart” AI’s in order to implement a better, smarter AI?
- Cognitive principal-agent problem: if an agent is maximizing a principal’s evaluation of “how good a job the agent did,” what properties of the principal will ensure good outcomes?
- Corrigibility as an attractor: what would it mean more formally for corrigible systems to define a broad basin of attraction towards acceptable outcomes, and is that likely to be the case?
- Robustness: can we achieve worst-case guarantees in learning systems?
- Semi-supervised RL: how few reward labels can we get away with for the problems we care about? How can we reason about this question in advance?
- Toy models of alignment: can we design any simple models that capture key aspects of the alignment problem but can be studied formally?
- Going for the throat: if we take plausible AI capabilities as given, can we design an aligned AI? Can we do it using more exotic resources like a hypercomputer?
- Benign induction: can we formally define an inductive process which generalizes reasonably quickly while avoiding the clearly “pathological” hypotheses that afflict solomonoff induction or logical induction? Alternatively, can we explain clearly why this won’t be a problem?

- New failure modes: can we identify any new alignment failures, that are plausible but haven't yet been discussed?
- Scalable transparency: can we better understand the internal behavior of sophisticated ML models, in a way that would help us prevent exotic failures like a treacherous turn and that would predictably scale up to very powerful models?
- Sampling IRL problems: can we sample from a distribution of agents such that (a) we "know" the values of those agents, and (b) the actual IRL problem for humans is in distribution?
- IRL over metacognition: can we learn human preferences over cognitive procedures, and use this to give convincing answers to "what would humans decide if they thought much longer / better?"
- Understanding consequentialism: can we develop any machinery for reasoning about how optimization and consequentialism appear and behave in our AI systems?
- Can we find invariants that help analyze AI systems built out of simpler parts? I'm especially interested in invariants of the form "not evil" rather than "aligned with human interests."
- Understanding universality and autopoiesis: can we understand what processes are "strong enough" that they can be said to have values and to converge upon deliberation? There is a big space of murky concepts here that seems important.
- How can we even define what the "right" behavior for an AI system is? I've usually thought about this in terms of "what deliberative processes do we endorse," but other approaches are also welcome.
- The easy goal inference problem: given unlimited time and perfect knowledge about human behavior, can we find any reasonable approximation to "what a human wants"?
- Messy evolution: can we reason well about evolutionary processes in which there is cultural development rather than selection on easily-isolated individuals? Will alignment be more difficult for these systems?
- Arguments for hardness: can we make more precise arguments about which alignment approaches won't work and why they are hard?

Renewal and certificates of impact

We'll be paying you for whatever you get done. Based on my experience as a researcher, I think that detailed oversight or roadmaps are often unhelpful for open-ended research projects.

Most of the time funders and researchers don't agree on any formal division of "credit.;" by default we'll do the same thing here. If that sounds good to you, no need to read further.

However, I am interested in using certificates of impact to be more precise about causal attribution and facilitate more effective prizes in the future. So after your funding, you'll have the option to treat it as an "advance" for certificates of impact, and then receive another round of funding. Here's how that works:

- You can ask me to evaluate the impact of your work, and I'll decide how many dollars I would have been willing to pay for that work in hindsight. You don't have to make a decision about renewal until you see this number.
- If you want to sell your impact and renew your funding, you decide how much of the causal credit you want to "sell" (up to 50%) and how much you want to "keep."
- You are paid: (how much I value the work) * (the fraction you want to sell). The original funding is now rolled forward.
- When doing the "altruistic accounting" for your work, you are encouraged to pretend that you were responsible only for whatever fraction of the impact you kept. Your altruistic impact is less than 100% of the project's impact because it diverted money which would have otherwise been used for other AI alignment projects. In general, the funder's impact and the contractor's impact should add up to 100% to avoid double-counting.

For example:

- Alice receives a 6 month contract for $60k.
- After 6 months, I evaluate her total output at $300k.
- She decides to renew funding for another 6 months, and to sell 30% of her original output. She receives $90k (= 30% * $300k).
- At that point, Alice is responsible for 70% of her impact over the initial 6 months and the funder is responsible for 30%, while her work over the following 6 months will be treated as traditional funding no formal allocation of credit.
- If instead I had evaluated her total output at $100k, Alice might decide not to renew. In this case, her work over the first 6 months would have no formal allocation of credit.

Previous rounds

I made a preliminary round of grants in October 2017. In that round I awarded:

- $20,000 to Peter Scheyer
- $10,000 to Chris Pasek
- $10,000 to Ryan Carey (not yet disbursed)