AI alignment topics

Paul Christiano

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?