My short-term agenda

Paul Christiano Follow
Mar 3, 2016

I think that the informed oversight problem is the biggest open problem with my recent proposal for aligned AI.

That said,

1. If we really established that informed oversight was the main missing ingredient, it would change our understanding of value alignment.
2. I think that other people are (justifiably) not yet convinced.
3. If informed oversight was the main missing ingredient, then our best bet may be to take an ad hoc approach to informed oversight even if there is no theoretically solid approach.
4. The informed oversight problem looks quite hard; I think it will be much easier to get traction on the other problems.

So my plan is to temporarily set aside the informed oversight problem and focus on the other difficulties. In rough order of priority:

1. Bootstrapping.
2. Robust prediction.
3. Robust active learning.
4. Efficient semi-supervised learning.

I don't think that any of these are straightforward, but I do feel pretty good about all of them. That view may change as I spend more time working on them.