Paul Christiano Follow
Dec 4, 2015

Consider a machine that does exactly what its user would tell it to do. If the user is a consequentialist, then so is the machine.

But building this machine does not introduce any *new* goals into the world at all. All of its consequentialism flows through the user's head —it merely amplifies the goal-directed reasoning that already happens there. There is no room to err in specifying its goals, because its goals are not specified.

This is the best case for act-based approaches to AI control.

But: this system may be optimizing internally, and is itself optimized.

We aim for *all* of this optimization to be a reflection and amplification of the user's preferences.

But: the user's reasoning is not perfect, and they may want AI to go beyond their capabilities.

We aim for humans to collaborate effectively with AI systems, forming teams that share human preferences and whose foresight exceeds the individual systems they are overseeing.

This project doesn't seem easy, but I feel optimistic.