AI “safety” vs “control” vs “alignment”

Paul Christiano Follow
Nov 19, 2016

Note: there has been some discussion recently about what term to use, and there are several legitimate complaints with control (for example, see Robby’s responses to this post). I’m tentatively moving towards the term “AI alignment” as a substitute for what is called “AI control” here.

I take these terms to describe a sequence of increasingly specific problems:

- **AI safety:** reducing risks posed by AI, especially powerful AI. Includes problems in misuse, robustness, reliability, security, privacy, and other areas. (Subsumes AI control.)
- **AI control**: ensuring that AI systems try to do the right thing, and in particular that they don’t competently pursue the wrong thing. I’ve argued that it’s roughly the same set of problems as AI security.
- **Value alignment**: understanding how to build AI systems that share human preferences/values, typically by learning them from humans. (An aspect of AI control.)

It would be great to hear if others use or understand these terms differently.

As you might guess from the title of this blog, my research is about AI control.