Clarifying “AI alignment”

Paul Christiano Follow
Apr 7, 2018 · 4 min read

When I say an AI A is *aligned with* an operator H, I mean:

> *A is trying to do what H wants it to do.*

The “alignment problem” is the problem of building powerful AI systems that are aligned with their operators.

This is significantly narrower than some other definitions of the alignment problem, so it seems important to clarify what I mean.

In particular, this is the problem of getting your AI to try to do the right thing, **not** the problem of figuring out which thing is right. An aligned AI would try to figure out which thing is right, and like a human it may or may not succeed.

Analogy

Consider a human assistant who is trying their hardest to do what H wants.

I’d say this assistant is aligned with H. If we build an AI that has an analogous relationship to H, then I’d say we’ve solved the alignment problem.

“Aligned” doesn’t mean “perfect:”

- They could misunderstand an instruction, or be wrong what H wants at a particular moment in time.
- They may not know everything about the world, and so fail to recognize that an action has a particular bad side effect.
- They may not know everything about H’s preferences, and so fail to recognize that a particular side effect is bad.
- They may build an unaligned AI (while attempting to build an aligned AI).

I use alignment as a statement about the *motives* of the assistant, not about their knowledge or ability. Improving their knowledge or ability will make them a better assistant—for example, an assistant who knows everything there is to know about H is less likely to be mistaken about what H wants—but it won't make them *more aligned*.

(For very low capabilities it becomes hard to talk about alignment. For example, if the assistant can't recognize or communicate with H, it may not be meaningful to ask whether they are aligned with H.)

Clarifications

- The definition is intended *de dicto* rather than *de re*. An aligned A is trying to "do what H wants it to do." Suppose A thinks that H likes apples, and so goes to the store to buy some apples, but H really prefers oranges. I'd call this behavior aligned because A is trying to do what H wants, even though the thing it is trying to do ("buy apples") turns out not to be what H wants: the *de re* interpretation is false but the *de dicto* interpretation is true.

- An aligned AI can make errors, including moral or psychological errors, and fixing those errors isn't part of my definition of alignment except insofar as it's part of getting the AI to "try to do what H wants" *de dicto*. This is a critical difference between my definition and some other common definitions. I think that using a broader definition (or the *de re* reading) would also be defensible, but I like it less because it includes many subproblems that I think (a) are much less urgent, (b) are likely to involve totally different techniques than the urgent part of alignment.

- An aligned AI would also be trying to do what H wants **with respect to clarifying H's preferences**. For example, it should decide whether to ask if H prefers apples or oranges, based on its best guesses about how important the decision is to H, how confident it is in its current guess, how annoying it would be to ask, *etc.* Of course, it may also make a mistake at the meta level—for example, it may not understand when it is OK to interrupt H, and therefore avoid asking questions that it would have been better to ask.

- This definition of "alignment" is extremely imprecise. I expect it to correspond to some more precise concept that cleaves reality at the joints. But that might not become clear, one way or the other, until we've made significant progress.

- One reason the definition is imprecise is that it's unclear how to apply the concepts of "intention," "incentive," or "motive" to an AI system. One naive approach would be to equate the incentives of an ML system with the objective it was optimized for, but this seems to be a mistake. For example, humans are optimized for reproductive fitness, but it is wrong to say that a human is incentivized to maximize reproductive fitness.

- "What H wants" is even more problematic than "trying." Clarifying what this expression means, and how to operationalize it in a way that could be used to inform an AI's behavior, is part of the alignment problem. Without additional clarity on this concept, we will not be able to build an AI that tries to do what H wants it to do.

Postscript on terminological history

I originally described this problem as part of "the AI control problem," following Nick Bostrom's usage in *Superintelligence*, and used "the alignment problem" to mean "understanding how to build AI systems that share human preferences/values" (which would include efforts to clarify human preferences/values).

I adopted the new terminology after some people expressed concern with "the control problem." There is also a slight difference in meaning: the control problem is about coping with the possibility that an AI would have different preferences from its operator. Alignment is a particular approach to that problem, namely avoiding the preference divergence altogether (so excluding techniques like "put the AI in a really secure box so it can't cause any trouble"). There currently seems to be a tentative consensus in favor of this approach to the control problem.

I don't have a strong view about whether "alignment" should refer to this problem or to something different. I do think that *some* term needs to refer to this problem, to separate it from other problems like "understanding what humans want," "solving philosophy," *etc.*