The alignment tax is the extra cost of ensuring that an AI system is aligned, relative to the cost of building an unaligned alternative. The term "tax" is used metaphorically here: in the AI safety literature, "alignment/safety tax" or "alignment cost" is meant to refer to all the additional costs of alignment — including increased developer time, extra compute, and decreased performance — and not only to the financial cost/tax required to build an aligned system.

In order to get a better idea of what the alignment tax is, consider two extreme possibilities.

- *No Tax*: The best case scenario, in which there is no cost to aligning the system, so there is no reason to deploy an AI that is not aligned.
- *Max Tax*: The worst case scenario, in which alignment is functionally impossible because an aligned system would take forever to develop, require infinite compute, or be completely useless. So you either deploy an unaligned system, or you don't get any benefit from AI systems at all.

In reality, we expect something in between these two scenarios to be the case.

Paul Christiano distinguishes two main approaches to dealing with the alignment tax:

1. Have the will to pay the tax, i.e., ensure that the relevant actors such as corporations and governments are willing to pay the extra costs to avoid deploying a system until it is aligned.
2. Reduce the tax by differentially advancing existing alignable algorithms or by making existing algorithms more alignable. This means, for any potentially unaligned algorithm, ensuring the additional cost for an aligned version of the algorithm is low enough that the developers would be willing to pay it.

Alternative phrasings

- Can safe AI design be competitive?

Related

- What is AI governance?

Notes for commenter/feedbacker(s):

- This definition is only here for feedback for now, will be moved to LW with a link to the tag when finished/approved.

Notes for self/ other author(s):

- Add links to - intent alignment when the question is live on site (or on LW)

Sources

- Yudkowsky, Eliezer (2017) Aligning an AGI adds significant development time
- Christiano, Paul (2020). Current work in AI alignment