Benign AI

Paul Christiano Follow
Nov 29, 2016 · 4 min read

Like it or not, humans are now in the business of building machines that optimize—we write code that trains computer vision systems, we design motion planning algorithms that control robots' behavior, and so on.

The goal of AI control is to ensure the results are optimized for our interests. This post introduces some definitions I find helpful for that project.

Talking about "our" interests is a little bit tricky, since different humans want different things. For concreteness I'll think about a particular system (e.g. Google Now running on your phone) and its "stakeholders:" its owners (you), designers (Google), and users (whoever is talking to the phone).

Something is ***malign*** if it is optimized for preferences that are incompatible with any combination of its stakeholders' preferences, i.e. such that over the long run using resources in accordance with the optimization's implicit preferences is not Pareto efficient for the stakeholders.

Something is ***benign*** if it is not malign.

I previously said "aligned" instead of "benign." But this isn't how other people use language, so I think that I need a new word. For example, I think that a toaster is benign, but apparently most people don't want to call a toaster "aligned."

Dynamics

Malignancy wants to spread. A malign computation will produce malign outputs. Malign outputs will be optimized to recruit other parts of the system to be malign, or to co-opt resources used by other parts of the system.

So it's especially important to understand how malignancy can enter into a system, and to either prevent it from entering or to limit its spread.

Today, malignancy is usually introduced by an adversary. For example, data received over the internet is likely to be malign. If you are designing a server, you need to be very careful to prevent the malignancy from spreading, and this turns out to be a surprisingly difficult problem.

Malignancy can also be introduced by certain algorithmic techniques. For example, gradient descent can produce policies that are highly optimized to have particular effects on the external world. Unless the optimization criterion reflects the user's preferences, those policies (and hence their outputs) will be malign.

For sufficiently weak optimization this isn't a problem, because the resulting policies aren't powerful enough to effectively spread or to compete with humans. But if we performed powerful enough optimization to produce human-level cognition, then we could have a problem.

Why the concept is useful

I think that benignity is useful as an invariant when designing and analyzing aligned AI systems.

On one end, it's easy to start with benign building blocks. User input is benign, since it's optimized by the user. The software we write is benign, as a special case of user input. Randomness is benign.

On the other end, powerful AI systems with benign outputs are necessarily aligned—their outputs are powerfully optimized, and not at all optimized for anything that we don't like, $\therefore$ their outputs are effectively optimized for our preferences.

So the game is to build a powerful AI without introducing any malignancy (and while coping with any malign data injected by an adversary). Each time we want to combine several ingredients, we can ask: assuming that the ingredients are benign, will the result be?

For example, in capability amplification the goal is to ensure that *if* the starting policy is benign, then the amplified policy is benign. In reward engineering the goal is to ensure that *if* the overseer is benign, then the learned policy is benign. And so on.

Why the concept is incomplete

I often use this notion of benignity, but it's definitely not completely satisfying.

For example, something can be optimized for different preferences to different extents; under realistic conditions I suspect that anything interesting will be weakly optimized in many different directions, and so *everything* is probably malign according to the all-or-nothing definition.

Realistically benignity needs to be understood quantitatively rather than bluntly and qualitatively. (Long ago Eliezer wrote about an approach to quantifying optimization, which I think is still at a "blunt qualitative" stage of elaboration.)

As another example, optimization is probably best understood with respect to some epistemic state, and we probably care about optimization with respect to *our* current beliefs.

There are many other subtleties along these lines, and we would need to resolve many of them in order to actually use benignity in any kind of precise analysis.

For now I think of benignity as a placeholder; I hope that eventually we'll be able to find a theoretically cleaner analog (and I'm counting on that cleaner analog to do a lot of heavy lifting).

I may be too optimistic, and benignity may not turn into a helpful concept for precise analysis. But I've already found the informal version to be useful for my own thinking.