When is unaligned AI morally valuable?

Paul Christiano Follow
May 2, 2018 · 12 min read

Suppose that AI systems built by humans spread throughout the universe and achieve their goals. I see two quite different reasons this outcome could be good:

1. Those AI systems are aligned with humans; their preferences are our preferences.
2. Those AI systems flourish on their own terms, and we are happy for them even though they have different preferences.

I spend most of my time thinking about option #1. But I think option #2 is a plausible plan B.

Understanding how happy we should be with an unaligned AI flourishing on its own terms, and especially *which* unaligned AIs we should be happy about, seems like a very important moral question.

I currently feel very uncertain about this question; if you forced me to guess, I'd estimate that option #2 allows us to recover 25% of the expected value that we lose by building unaligned AI. But after more thinking, that number could go down to 0% or up to >90%.

Definition

In this post I'll say that an AI is a *good successor* if I believe that building such an AI and "handing it the keys" is a reasonable thing to do with the universe. Concretely, I'll say an AI is a good successor if I'd prefer give it control of the world than accept a gamble where we have a 10% chance of extinction and a 90% chance of building an aligned AI.

In this post I'll think mostly about what happens with the rest of the universe, rather than what happens to us here on Earth. I'm wondering whether we would appreciate what our successors do with all of the other stars and galaxies—will we be happy with how they use the universe's resources?

Note that a competent aligned AI is a good successor, because "handing it the keys" doesn't actually amount to giving up any control over the universe. In this post I'm wondering which unaligned AIs are good successors.

Preface: in favor of alignment

I believe that building an aligned AI is by far the most likely way to achieve a good outcome. An aligned AI allows us to continue refining our own views about what kind of life we want to exist and what kind of world we want to create—there is no indication that we are going to have satisfactory answers to these questions prior to the time when we build AI.

I don't think this is parochial. Once we understand what makes life worth living, we can fill the universe with an astronomical diversity of awesome experiences. To the extent that's the right answer, it's something I expect us to embrace much more as we become wiser.

And I think that further reflection is a really good idea. There is no law that the universe tends towards universal love and goodness, that greater intelligence implies greater moral value. Goodness is something we have to work for. It might be that the AI we would have built anyway will be good, or it might not be, and it's our responsibility to figure it out.

I am a bit scared of this topic because it seems to give people a license to hope for the best without any real justification. Because we only get to build AI once, reality isn't going to have an opportunity to intervene on people's happy hopes.

Clarification: Being good vs. wanting good

We should distinguish two properties an AI might have:

- Having preferences whose satisfaction we regard as morally desirable.
- Being a moral patient, e.g. being able to suffer in a morally relevant way.

These are **not** the same. They may be related, but they are related in an extremely complex and subtle way. From the perspective of the long-run future, we mostly care about the first property.

As compassionate people, we don't want to mistreat a conscious AI. I'm worried that compassionate people will confuse the two issues—in arguing enthusiastically for the claim "we should care about the welfare of AI" they will also implicitly argue for the claim "we should be happy with whatever the AI chooses to do." Those aren't the same.

It's also worth clarifying that both sides of this discussion can want the universe to be filled with morally valuable AI eventually, this isn't a matter of carbon chauvinists vs. AI sympathizers. The question is just about how we choose what kind of AI we build—do we hand things off to whatever kind of AI we can build today, or do we retain the option to reflect?

Do all AIs deserve our sympathy?

Intuitions and an analogy

Many people have a strong intuition that we should be happy for our AI descendants, whatever they choose to do. They grant the *possibility* of pathological preferences like paperclip-maximization, and agree that turning over the universe to a paperclip-maximizer would be a problem, but don't believe it's realistic for an AI to have such uninteresting preferences.

I disagree. I think this intuition comes from analogizing AI to the children we raise, but that it would be just as accurate to compare AI to the corporations we create. Optimists imagine our automated children spreading throughout the universe and doing their weird-AI-analog of art; but it's just as realistic to imagine automated PepsiCo spreading throughout the universe and doing its weird-AI-analog of maximizing profit.

It might be the case that PepsiCo maximizing profit (or some inscrutable lost-purpose analog of profit) is intrinsically morally valuable. But it's certainly not obvious.

Or it might be the case that we would never produce an AI like a corporation in order to do useful work. But looking at the world around us today that's *certainly* not obvious.

Neither of those analogies is remotely accurate. Whether we should be happy about AI "flourishing" is a really complicated question about AI and about morality, and we can't resolve it with a one-line political slogan or crude analogy.

On risks of sympathy

I think that too much sympathy for AI is a real risk. This problem is going to made particularly serious because we will (soon?) be able to make AI systems which are optimized to be sympathetic. If we are indiscriminately sympathetic towards whatever kind of AI is able to look sympathetic, then we can't steer towards the kind of AI that actually deserve our sympathy. It's very easy to imagine the world where we've built a PepsiCo-like AI, but one which is much better than humans at seeming human, and where people who suggest otherwise look like moral monsters.

I acknowledge that the reverse is also a risk: humans are entirely able to be terrible to creatures that deserve our sympathy. I believe the solution to that problem is to actually think about what the nature of the AI we build, and especially to behave compassionately in light of uncertainty about the suffering we might cause and whether or not it is morally relevant. Not to take an indiscriminate pro-AI stand that hands the universe over to the automated PepsiCo.

Do any AIs deserve our sympathy?

(Warning: lots of weird stuff.)

In the AI alignment community, I often encounter the reverse view: that *no* unaligned AI is a good successor.

In this section I'll argue that there are at least some unaligned AIs that would be good successors. If we accept that there are *any* good successors, I think that there are probably lots of good successors, and figuring out the boundary is an important problem.

(To repeat: I think we should try to avoid handing off the universe to any unaligned AI, even if we think it is probably good, because we'd prefer retain the ability to think more about the decision and figure what we really want. See the conclusion.)

Commonsense morality and the golden rule

I find the golden rule very compelling. This isn't just because of repeated interaction and game theory: I'm strongly inclined to alleviate suffering even if the beneficiaries live in abject poverty (or factory farms) and have little to offer me in return. I'm motivated to

help largely because that's what I would have wanted them to do if our situations were reversed.

Personally, I have similar intuitions about aliens (though I rarely have the opportunity to help aliens). I'd be hesitant about the people of Earth screwing over the people of Alpha Centauri for many of the same reasons I'd be uncomfortable with the people of one country screwing over the people of another. While the situation is quite confusing I feel like compassion for aliens is a plausible "commonsense" position.

If it is difficult to align AI, then our relationship with an unaligned AI may be similar to our relationship with aliens. In some sense we have all of the power, because we got here first. But if we try to leverage that power, by not building any unaligned AI, then we might run a significant risk of extinction or of building an AI that no one would be happy with. A "good cosmic citizen" might prefer to hand off control to an unaligned and utterly alien AI, than to gamble on the alternative.

If the situation were totally symmetrical—if we believed the AI was from *exactly* the same distribution over possible civilizations that we are from—then I would find this intuitive argument extremely compelling.

In reality, there are almost certainly differences, so the situation is very confusing.

A weirder argument with simulations

The last argument gave a kind of common-sense argument for being nice to some aliens. The rest of this post is going to be pretty crazy.

Let's consider a particular (implausible) strategy for building an AI:

- Start with a simulation of Earth.
- Keep waiting/restarting until evolution produces human-level intelligence, civilization, *etc.*
- Once the civilization is *slightly below* our stage of maturity, show them the real world and hand them the keys.
- (This only makes sense if the simulated civilization is much more powerful than us, and faces lower existential risk. That seems likely to me. For example, the resulting AIs would likely

think *much* faster than us, and have a much larger effective population; they would be very robust to ecological disaster, and would face a qualitatively easier version of the AI alignment problem.)

Suppose that *every* civilization followed this strategy. Then we'd simply be doing a kind of interstellar shuffle, where each civilization abandons their home and gets a new one inside of some alien simulation. It seems much better for everyone to shuffle than to accept a 10% chance of extinction.

Incentivizing cooperation

The obvious problem with this plan is that not everyone will follow it. So it's not really a shuffle: nice civilizations give up their planet, while mean civilizations keep their original planet *and* get a new one. So this strategy involves a net transfer of resources from nice people to mean people: some moral perspectives would be OK with that, but many would not.

This obvious problem has an obvious solution: since you are simulating the target civilization, you can run extensive tests to see if they seem nice—i.e. if they are the kind of civilization that is willing to give an alien simulation control rather than risk extinction—and only let them take over if they are.

This guarantees that the nice civilizations shuffle around between worlds, while the mean civilizations take their chances on their own, which seems great.

More caveats and details

This procedure might look really expensive—you need to simulate a whole civilization, nearly as large as your own civilization, with computers nearly as large as your computers. But in fact it doesn't require literally simulating the civilization up until the moment when they are building AI— you could use cheaper mechanisms to try to guess whether they were going to be nice a little bit in advance, e.g. by simulating large numbers of individuals or groups making particularly relevant decisions. If you were simulating humans, you could imagine predicting what the modern world would do without ever actually running a population of >100,000.

If only 10% of intelligent civilizations decide to accept this trade, then running the simulation is 10x as expensive (since you need to try 10 times). Other than that, I think that the calculation doesn't actually

depend very much on what fraction of civilizations take this kind of deal.

Another problem is that people may prefer continue existing in their own universe than in some weird alien simulation, so the "shuffle" may itself be a moral catastrophe that we should try to avoid. I'm pretty skeptical of this:

- You could always later perform an acausal trade to "go home," i.e. to swap back with the aliens who took over your civilization (by simulating each other and passing control back to the original civilization if their simulated copy does likewise).
- In practice the universe is very big, and the part of our preferences that cares about "home" seems easily satiable. There is no real need for the new residents of our world to kill us, and I think that we'd be perfectly happy to get just one galaxy while the new residents get everything else. (Given that we are getting a whole universe worth of resources somewhere else.)

Another problem is that this is a hideously intractable way to make an AI. More on that two sections from now.

Another problem is that this is completely insane. I don't really have any defense, if you aren't tolerant of insanity you should probably just turn back now.

Decision theory

The above argument about trade / swapping places makes sense from a UDT perspective. But I think a similar argument should be persuasive even to a causal decision theorist.

Roughly speaking, you don't have much reason to think that you are on the outside, considering whether to instantiate some aliens, rather than on the inside, being evaluated for kindness. If you are on the outside, instantiating aliens may be expensive. But if you are on the inside, trying to instantiate aliens lets you escape the simulation.

So the cost-benefit analysis for being nice is actually pretty attractive, and is likely to be a better deal than a 10% risk of extinction.

(Though this argument depends on how accurately the simulators are able to gauge our intentions, and whether it is possible to look nice but ultimately defect.)

How sensitive is moral value to the details of the aliens?

If an AI is from *exactly* the same distribution that we are, I think it's particularly likely that they are a good successor.

Intuitively, I feel like goodness probably doesn't depend on incredibly detailed facts about our civilization. For example, suppose that the planets in a simulation are 10% smaller, on average, than the planets in the real world. Does that decrease the moral value of life from that simulation? What if they are 10% larger?

What if we can't afford to wait until evolution produces intelligence by chance, so we choose some of the "randomness" to be particularly conducive to life? Does that make all the difference? What if we simulate a smaller population than evolution over a larger number of generations?

Overall I don't have very strong intuitions about these questions and the domain is confusing. But my weak intuition is that none of these things should make a big moral difference.

One caveat is that in order to assess whether a civilization is "nice," you need to see what they would do *under realistic conditions*, i.e. conditions from the same distribution that the "basement" civilizations are operating under. This doesn't necessarily mean that they need to evolve in a physically plausible way though, just that they *think* they evolved naturally. To test niceness we could evolve life, then put it down in a world like ours (with a plausible-looking evolutionary record, a plausible sky, *etc.*)

The decision-theoretic / simulation argument seems more sensitive to details than the commonsense morality argument. But even for the decision-theoretic argument, as long as we create a historical record convincing enough to fool the simulated people, the same basic analysis seems to apply. After all, how do we know that *our* history and sky aren't fake? Overall the decision-theoretic analysis gets really weird and complicated and I'm very unsure what the right answer is.

(Note that this argument is very fundamentally different from using decision theory to constrain the behavior of an AI—this is using decision theory to guide our *own* behavior.)

Conclusion

Even if we knew how to build an unaligned AI that is *probably* a good successor, I still think we should strongly prefer to build aligned AGI. The basic reason is option value: if we build an aligned AGI, we keep all of our options open, and can spend more time thinking before making any irreversible decision.

So why even think about this stuff?

If building aligned AI turns out to be difficult, I think that building an unaligned good successor is a plausible Plan B. The total amount of effort that has been invested in understanding which AIs make good successors is very small, even relative to the amount of effort that has been invested in understanding alignment. Moreover, it's a separate problem that may independently turn out to be much easier or harder.

I currently believe:

- There are definitely some AIs that aren't good successors. It's probably the case that many AIs aren't good successors (but are instead like PepsiCo)
- There are very likely to be some AIs that are good successors but are very hard to build (like the detailed simulation of a world-just-like-Earth)
- It's plausible that there are good successors that are easy to build.
- We'd likely have a *much* better understanding of this issue if we put some quality time into thinking about it. Such understanding has a really high expected value.

Overall, I think the question "which AIs are good successors?" is both neglected and time-sensitive, and is my best guess for the highest impact question in moral philosophy right now.