Link-sharing is on, and the link has been shared in channels with about ~50 unique people in them. If you want to leave private comments, please make a copy and share it with me.

Implications of ECL

"ECL" is short for "evidential cooperation in large worlds". It's an idea that was originally introduced in Oesterheld (2017) (under the name of "multiverse-wide superrationality"). This post will explore implications of ECL, but it won't explain the idea itself. If you haven't encountered it before, you can read the paper linked above or this summary written by Lukas Gloor.[1]

This post briefly lists what I see as the main plausible decision-relevant implications of ECL. In this post, I will not justify these implications in much depth. (Though for some implications, I follow-up with more analysis and justification in other posts, see more.) Instead, I will lean on the principle that ECL recommends that we (and other ECL-sympathetic actors) act to benefit the values of people whose decisions might correlate with our decisions.

As described in this appendix, this relies on you and others having particular kinds of values. For one, I assume that you care about what happens outside our light cone. But more strongly, I'm looking at values with the following property: If you could have a sufficiently large impact outside our lightcone, then the value of taking different actions would be dominated by the impact that those actions had outside our lightcone. I'll refer to this as "universe-wide values". Even if *all* your values aren't universe-wide, I suspect that the implications will still be relevant to you if you have *some* universe-wide values.

This is speculative stuff, and I'm not particularly confident that I will have gotten any particular claim right.

Summary

For at least two reasons, future actors will be in a better position to act on ECL than we are. Firstly, they will know a lot more about what other value-systems are out there. Secondly, they will be facing immediate decisions about what to do with the universe, which should be informed by what other civilizations would prefer.[2] This suggests that it could be important for us to Affect whether (and how) future actors do ECL. This can be decomposed into two sub-points that deserve separate attention: how we might be able to affect Futures with aligned AI, and how we might be able to affect Futures with misaligned AI.

[1] For even more references, see all the content gathered on this page, and more recently, this post written by Paul Christiano and this paper by Johannes Treutlein.
[2] Whereas today, we can focus on handing-off the future to a broadly competent and healthy civilization, and trust decisions about what to do with the future to them.

But separately from influencing future actors, ECL also changes our own priorities, today. In particular, ECL suggests that we should care more about other actors' universe-wide values. When evaluating these implications, we can look separately at three different classes of actors and their values. I'll separately consider how ECL suggests that we should…

- Care more about *other humans'* universe-wide values.[3]
    - I think the most important implication of this is that Upside- and downside-focused longtermists should care more about each others' values.
- Care more about *evolved aliens'* universe-wide values.
    - I think the most important implication of this is that we plausibly should care more about influencing how AI could benefit/harm alien civilizations.
- Care more about *misaligned AIs'* universe-wide values.[4]
    - I don't think this significantly reduces the value of working on alignment.
    - But it suggests that it could be valuable to build AI so that *if* it ends up misaligned, it has certain other desirable inclinations and values. This topic, of positively influencing misaligned AI in order to cooperate with distant misaligned AI, is very gnarly, and it's difficult to tell what sort of changes would be net-positive vs. net-negative.

(For more details on the split between humans/evolved-aliens/misaligned-AI and why I chose it, see this appendix.)

[3] When I discuss how we should "care more about other humans' universe-wide values", I exclusively refer to universe-wide values held by humans on our current planet Earth, as opposed to values that might be held by distant human-like species. But the reason to benefit such values is to generate evidence that other people benefit our values on distant planets (not just here, on planet Earth). So why focus specifically on humans' values? The reason is that we are more confident that some people treasure them, and it's easy to benefit them via supporting humans who support them. For more, see here.

[4] "Misaligned AI" refers to AI whose values are very different from what was intended by the evolved species that first created them. If a distant species has very different values from us, and successfully aligns AI systems that they create, I wouldn't count those as "misaligned AIs".

Implications of ECL **1**
Summary 1
Affect whether (and how) future actors do ECL 3
Futures with aligned AI 3
Futures with misaligned AI 4
How us doing ECL affects our priorities 4
Care more about other humans' universe-wide values 4
It matters less which universe-wide values control future resources (seems minor in practice?) 4
Upside- and downside-focused longtermists should care more about each others' values 6
Care more about evolved aliens' universe-wide values 7
Minor: Prioritize non-AI extinction risk less highly 8
Influence how AI benefits/harms alien civilizations' values 8
Care more about misaligned AIs' universe-wide values 9
Minor: Prioritize AI takeover risk less highly 9
Positively influence misaligned AI 10
More 10
Appendices **10**
What values do you need for this to be relevant? **10**
More details on the split between humans, evolved species, and misaligned AI 11
Why distinguish humans from aliens? 11
Why distinguish evolved aliens from misaligned AIs? 12

Affect whether (and how) future actors do ECL

Futures with aligned AI

If we take ECL seriously,[5] I think it's really important that humanity *eventually* understands these topics deeply, and can make wise decisions about them. But for most questions about what humanity should *eventually* do, I'm inclined to defer them to the future. I'm interested in whether there's anything that *urgently* needs to be done.

One way to affect things is to increase the probability that humanity ends up building a healthy and philosophically competent civilization. (But we already knew that was important.)

There might also be ways in which humanity could irreversibly mess up in the near-term that are unique to decision theory. For example, people could make unwise commitments if they perceive

[5] Or any other kind of acausal effects.

themselves to be in commitment races.[6] Or there might be ways in which people could learn too much information, too early. (We don't currently have any formalized decision theories that *can't* be harmed by learning information. For example, people who use evidential decision theory can only influence things that they haven't yet learned about — which means that information can make them lose power.) (C.f. this post.) It's possible that careful thinking could reduce such risks.[7] (For example, perhaps it would be good to prevent early AI systems from thinking about these topics until they and we are more competent.)

How is ECL relevant for this? Broadly, it seems like ECL is an important part of the puzzle for what various decision theories recommend. So learning more about ECL seems like it could help clarify the picture, here, and clarify what intervention points exist. (This also applies to futures with misaligned AI.)

(For discussion in Oesterheld (2017), see section 4.5 on researching and promoting ECL and 4.6.3 on making AI act according to ECL.)

Futures with misaligned AI

Affecting how misaligned AI does ECL is also an intervention point.

I think ECL could play a couple of different roles, here:

- Firstly, ECL-sympathetic AI systems might treat *us* better (e.g. by giving humanity a solar-system-sized utopia instead of killing us).
    - In order for ECL to recommend this, there needs to be some distant actors that care about us. I.e., someone would need to have universe-wide values that specifically values the preservation of distant pre-AGI civilizations over other things that could be done with those resources.
- Secondly, ECL-sympathetic AI systems might trade (and avoid conflict) with distant civilizations, thereby benefiting those civilizations.
    - This is intrinsically good if we intrinsically care about those distant civilizations' values.
    - In addition, it's plausible that ECL recommends us to care about benefits that accrue to distant civilizations' whose values we don't intrinsically care about. This is discussed below, in Influence how AI benefits/harms alien civilizations.
    - Such trade could also benefit *the misaligned AI system's own values*, and ECL might give us reason to care about those values. This is more complicated. I discuss it more in Positively influence misaligned AI.

[6] Premature commitments are often a gamble that might gain *you* a better bargaining position while carrying a risk of *everyone* getting a lower payoff. Since that's quite uncooperative, it seems plausible that ECL could discourage premature commitments. So this might be a reason to spread knowledge about ECL.

[7] Though also possible that *un*careful thinking could increase them — given that they are by-their-nature caused by humanity making errors in what order they learn about and commit to doing certain things.

How *us doing* ECL affects our priorities

Care more about other humans' universe-wide values

It matters less which universe-wide values control future resources (seems minor in practice?)

Let's temporarily assume that humanity will avoid both near-term extinction and AI takeover. Even then, the value of the future could depend a lot on *which human values* will be empowered to decide what's to be done with all the matter, space, and time in our lightcone.

If someone had an opportunity to influence this (e.g. by promoting certain values), ECL would generally be positive on empowering universe-wide values (that are compatible with good decision-theoretic reasoning), since for any such values:

- You might correlate with distant people who hold such values, in which case ECL gives you reason to benefit them.
- If such values maintain power into the long-term future, and our future civilization ends up deciding that ECL (or something similar) works, then ECL will motivate them to benefit other universe-wide values. (At least insofar as there are gains from trade to be had.)

If you were previously concerned about promoting *any particular* universe-wide values, this means that you should now be somewhat less fussed about promoting those values in particular, as opposed to any other universe-wide values. In struggles for influence that are mainly a struggle about universe-wide values, you should care less about who wins.

(This is related to discussion about moral advocacy in section 4.2 of Oesterheld (2017); especially 4.2.7.)

Here's a slightly more worked-out gesture at why ECL would recommend this.

- Let's say that you're a supporter of faction A, in a struggle for influence against faction B. You can decide to either invest in the 0-sum struggle for influence, or you can decide to invest in something that you both value (e.g. reducing uncontroversial x-risks or s-risks).
- If support for faction B is compatible with good decision-theoretic reasoning, then on some distant planet, there will probably be supporters of faction B who are in an analogous but reversed situation to you (in a struggle for influence against faction A) who are thinking about this decision in a similar way.
- If you decide to support the common good instead of faction A, then faction A's expected influence will decrease a bit on your planet. But your choice to do so is evidence that the distant supporters of faction B also will support the public good (instead of faction B) on their planet, which will lead faction A's expected influence to increase a bit (and also lead to positive effects from the support of the public good).
- So ECL provides a reason to invest less resources in the 0-sum fight and instead care more about public goods.

(In order to work out *how* much less you'd want to invest in the 0-sum fight, you'd want to think about the ratio between "how much evidence am I providing that supporters of faction A will invest in the public good" to "how much evidence am I providing that supports of faction B will invest in the public good".[8] I'm only illustrating the directional argument, here.)

While I believe the ECL argument works here, it doesn't currently seem very decision-relevant to me. Competitions that could be important for the future (e.g. competition between AI labs or between US and China) don't seem well-characterized as conflicts between universe-wide value-systems. At least my personal opinions about them are mostly about who's more/less likely to cause an (uncontroversial) x-risk along the way; and perhaps about who's more/less likely to help create a society that adopts reasonably impartial values and become sufficiently philosophically sophisticated to enact the best version of them.

That said, for someone who was previously obsessed with boosting a *particular* value-system (e.g. spreading hedonistic utilitarianism, or personally acquiring power for impartial ends), I think ECL should weaken/change that motivation to be somewhat more inclusive of other universe-wide values.

Upside- and downside-focused longtermists should care more about each others' values

(Terms are defined as here: Upside-focused values are values that *in our empirical situation* recommend focusing on bringing about lots of positive values. Downside-focused values are values that *in our empirical situation* recommend working on interventions that make bad things less likely, typically reducing suffering.)

If we look beyond struggles for influence and resources, and instead look for any groups of humans who have different *universe-wide* values, and where this leads to different real-world priorities, the two groups that stand out are upside-focused and downside-focused longtermists. For these groups, we also *have actual examples* of both upside- and downside-focused people thinking about ECL-considerations in a similar way. Which makes the ECL-argument more robust.

It seems good for people to know about and bear this in mind. For example, it means that upside- and downside-focused people should:

- be inclined to take high-leverage opportunities to help each other,
- decide what to work on somewhat less on the basis of values and somewhat more on the basis of comparative advantage,
- avoid actions that would benefit their own values at considerable cost to the others' values.

[8] And ideally, you would also think about other opportunities that faction A and faction B would have of benefiting each other, since you might also be providing evidence about those. Even more ideally, you might think about possible gains from trades that involve even more factions.

As usual, the ECL-argument here is: If you choose to take any of these actions, then that's evidence that distant people with *the other* value-system will choose to take analogous actions to benefit *your* favorite value-system.

How strong is this effect? I'm not sure. What follows is a few paragraphs of speculation. (Flag: These paragraphs rely more on pre-existing knowledge about ECL than the rest of the post.)

Ultimately, it depends on the degree to which humans correlate relatively more with the decisions of people with shared values vs. different values, on this type of decision.

I.e., the question is: If someone with mostly upside-focused values decides to do something that benefits downside-focused values, how much evidence is this that (i) distant upside-focused people will help out people with downside-focused values, vs. (ii) distant downside-focused people will help out people with upside-focused values. (From the perspective of the person who makes the decision.)

If it's similarly much evidence for both propositions, then upside-focused and downside-focused people should be similarly inclined to benefit each others' values as to benefit their own values.[9]

Here's an argument in favor of this: Regardless of whether you have upside-focused or downside-focused values, the ECL argument (that you should care about the others' values) is highly similar. So it seems like there's no large dependence on what values you have. Accordingly, it seems like your decision should be equally much evidence for how other people act, regardless of what values they have.

Here's a counter-argument: When you're making any particular decision, perhaps you are getting disproportionately much evidence about how actors that are especially similar to you tend to feel about ECL-based cooperation-arguments in especially similar situations. (After all: Most of your evidence about how likely people are to act on ECL-arguments *in general* will come from observations about what decisions *other* people make.) And perhaps an important component of "especially similar" is that those actors would share your values.[10]

(For some more of my speculation, including some counter-arguments to that last paragraph, see the post Can we influence distant AIs' choices? — LF Jul 2023, which discusses a similar question but with regards to correlating with *AI* instead of with *humans*. A relatively less likely

[9] Though the total effort that goes to each should perhaps still be allocated based on the number of people who support each set of values and who are sympathetic to ECL. Potentially adjusted by speculation about whether either set of values is underrepresented (among ECL-sympathizers) on Earth compared to the universe-at-large, in which case we should prioritize that set of values higher.

[10] It will be *the most* evidence for the actions of people in *exactly* my position. But this is not where most of my acausal influence will come from, since even a small amount of evidence across a sufficiently larger number of actors will weigh higher. The hypothesis that I'm putting forward here is that there might be some fairly broad class of actors which still share some key similarities with you, whose decisions your decisions provide more evidence about. And that your values might be (or be correlated with) one of the key similarities.

proposition. Nevertheless — similar considerations come up. See also section 3.1 in Oesterheld (2017) on orthogonality of rationality and values.)

Overall, I feel quite uncertain here. This uncertainty corresponds to a sense that my actions are somewhat less evidence for the decisions of people who don't share my values, but not a huge amount less. Summing over my uncertainty, I feel like my decisions are ≥10% as much evidence for the decisions of people who don't share my value (as they are evidence for the decisions of people who share my values) — which would imply that I should care ≥10% as much about their values as I care about my own.[11]

Care more about evolved aliens' universe-wide values

ECL also recommends caring more about alien civilizations. Here are two different implications of this.

Minor: Prioritize non-AI extinction risk less highly

A minor consequence of this is: You might want to prioritize non-AI extinction risk slightly *less* highly than before. Because if Earth doesn't colonize the universe, some of that space will (in expectation) get colonized by alien civilizations instead, to their benefit.

If we were to trade like-for-like, the story would be: If we prioritize non-AI extinction risk slightly less highly (and put higher priority on making sure that space colonization is good if it does happen), then that's evidence that distant aliens also prioritize non-AI extinction risk slightly less highly. If this leads to their extinction, and their neighbors share our values, then civilizations with our values will recover some of that empty space.[12]

I think this effect seems minor (unlikely to make non-AI extinction less than half as useful as you previously thought). Because:

- Aliens are (probably) not common enough to take over all space that we would have missed. I think the relevant comparison is between "space we could get to *before aliens*" (i.e. the total amount of space that humanity would get access to if space is defense-dominant) vs. "space that *only* we could get to" (i.e. the space that humanity would get access to without excluding any aliens from it, such that we would want to get there even if we cared just as much about alien's values as our own values). My old estimate is that the latter is ~⅓ times as large as the former. This suggests that, even if ECL made us care just as much about alien values as our own values, we would still care ⅓ as much about colonizing space.

[11] Though I am personally somewhat sympathetic to both upside- and downside-focused values, so this doesn't have a big impact on my all-things-considered view.

[12] Even if the aliens who went extinct shared our values, their choice to prioritize non-AI extinction risk less could still have been net-positive ex-ante. For example, they might have reallocated resources in a way that reduced AI takeover risk by 0.1% and increased non-AI extinction risk by 0.1001%. The added 0.0001% of x-risk might have been worth the benefit of leaving behind empty space rather than AI-controlled space in 0.1% of worlds.

- ECL doesn't recommend us to care *as* much about alien values as we care about human values.[13] I would be surprised if ECL recommended that we prioritize random alien values more than half as much as our own, which suggests that even if aliens were guaranteed to colonize space in our place, at least ½ of value would be lost from our failure to do so.

Also, as I discuss below, ECL might similarly motivate us to prioritize AI takeover less highly. Since this is the most salient alternative priority to "non-AI extinction risk" on a longtermist view, they partly cancel out.

Influence how AI benefits/harms alien civilizations' values

A different way in which we could benefit aliens is to increase the probability that Earth-originating intelligence benefits them (if Earth-originating intelligence doesn't go extinct). This could apply to either aliens that we physically meet in space, or to distant aliens that we can't causally interact with.

I typically think about such interventions as a special kind of "cooperative AI"-intervention — increasing the probability that AIs are inclined to seek out win-win opportunities with other value systems. See this post for more discussion of this. The brief summary is: ECL could plausibly increase the value of interventions that aim to make misaligned AI treat alien civilizations better by ~1.5-10x.

Care more about misaligned AIs' universe-wide values

More speculatively, ECL might recommend that we care more about the universe-wide values of distant misaligned AIs.[14] Why is this more speculative? In order for ECL to give us reason to benefit AIs, we would have to be similar enough to those AIs that our decisions have some acausal influence on their decisions. If we assume evidential decision theory, this means that our decision needs to give some evidence for what distant misaligned AIs choose to do. And intuitively, it seems less likely that our decisions provide evidence about distant misaligned AI's actions than that it provides evidence about the actions of distant aliens. (Since AIs' minds probably differ more from ours, and since the decision-situations they are likely to find themselves in differ more from ours.)

I feel uncertain about whether ECL says we should care more about the values of distant AIs. If it did, here are two conclusions.

Minor: Prioritize AI takeover risk less highly

One potential implication could be that we should prioritize AI takeover risk slightly less highly: Because although it would be bad for misaligned AI to seize power, it would at least be slightly good

[13] In particular, ECL suggests that we should discount benefits to aliens insofar as they on average correlate less strongly with us than the average civilizations-with-our-values do. (When making relevant decisions.)

[14] "Misaligned AI" refers to AI whose values are very different from what was intended by the evolved species that first created them. If a distant species has very different values from us, and successfully aligns AI systems that they create, I wouldn't count those as "misaligned AIs".

that the AI gets to implement its own values, as long as those values were endorsed by many distant ECL-sympathetic AIs. (Since us benefiting their values, in this way, would be evidence that they make decisions to benefit our values — or at least that's the hypothetical we're considering.)

However, this effect seems minor (unlikely to make misalignment reduction less than half as useful as you previously thought). Because:

- As mentioned above, we might not be similar enough to AIs for the ECL argument to work. And even if we're similar enough to have some acausal influence on them, ECL doesn't recommend us to care *as* much about AI values as we care about human values.[15] I would be surprised if ECL recommended that we prioritized random AI values more than half as much as our own.
- This is mainly an argument to slightly deprioritize AI takeover *by ECL-sympathetic AI systems with universe-wide values*. But AIs might not have universe-wide values, and might not be ECL-sympathetic, by default.

Also, as discussed above, ECL might similarly motivate us to prioritize *non-AI* extinction less highly. (Which is the most salient alternative priority to misalignment risk, on a longtermist view.)

Positively influence misaligned AI

A different way in which we could benefit distant misaligned AIs' universe-wide values is to adjust how we build AI so that *if* humanity ends up building AI that is misaligned with our own values, then it's more likely to successfully optimize for values that distant misaligned AIs would approve of. Unfortunately, it seems very difficult to work out what sort of changes would be good and what sort of changes would be bad, here.

For more writing on this, see here.

More

I'm following up this post with two other posts:

- ECL and benefitting distant civilizations — LF Jul 2023, for more on how ECL affects the value of influencing how AI might benefit/harm distant alien civilizations.
- summary and links for ECL with AI — LF Jul 2023, digging into how we could Positively influence misaligned AI, and whether ECL recommends that.

[15] In particular, ECL suggests that we should discount benefits to AI insofar as they correlate less strongly with us than actors-with-our-values do.

Appendices

What values do you need for this to be relevant?

I'd say this doc is roughly: advice to people whose values are such that, *if* they were to grant that their actions acausally affected an enormous number of worlds quite different from their own, they would say that a large majority of the impact they cared about was impact on those distant worlds.

Importantly, it's also advice to people who endorse some type of moral uncertainty or pluralism, and have *components* of their values that behave like that. Then it's advice for what that value-component should advocate and bargain for. (I think this is probably a more realistic account of most humans' values.)

(Though one of the many places where I haven't thought about the details is: If you are trying to acausally influence agents with universe-wide values, does it pose any extra troubles if you yourself only have partially universe-wide values and do some messy compromise thing?)

I'll use "universe-wide" values as a shorthand for these types of values. ("Multiverse-wide" would also be fine terminology — but I think caring about a spatially large universe is sufficient.)

(For previous discussion of what values are necessary for ECL, see section 3.2 in Oesterheld (2017).)

More details on the split between humans, evolved species, and misaligned AI

Above, I separately consider how ECL suggests that we should care more about:

- other humans' universe-wide values,
- evolved aliens' universe-wide values,
- misaligned AIs' universe-wide values.

This raises two questions:

- Why the distinction between "other humans' universe-wide values" and "evolved aliens' universe-wide value"?
- Why the distinction between "evolved aliens' universe-wide values" and "misaligned AIs' universe-wide values?

Why distinguish humans from aliens?

When I talk about benefiting other humans' universe-wide values, I don't mean to imply that we're acausally cooperating with just the local humans on our own planet Earth. I think almost all the benefits come via evidence that very distant actors behave more cooperatively. Such actors could be either quite similar to humans or quite unlike humans (in at least some ways).

So why talk specifically about the universe-wide values of "other humans", rather than the broader group of aliens?

The answer is that universe-wide values held by other humans has a number of unique properties:

- For any universe-wide values held by humans, we have *empirical support* that evolved species sometimes grow to treasure those values.
- Even stronger, we have empirical support that *minds very similar to our own* can grow to treasure those values, which strengthens the case for high correlations, and thereby the case for ECL-based cooperation.
- Universe-wide values held by humans can conveniently be benefitted *via* the humans that support them For example, by:
    - supporting the humans that hold them.
    - avoiding conflicts with humans that hold them.
    - listening to the advice of humans that hold them.

This is quite different from non-human values, where we have to resort to more basic guesses about their preferences, like:

- Aliens probably value having access to more space over having access to less space.
- Aliens probably prefer to interact with other actors who are cooperative rather than conflict-prone.

Why distinguish evolved aliens from misaligned AIs?

First, a terminological note: "Misaligned AI" refers to AI whose values are very different from what was intended by the evolved species that first created them. If a distant species has very different values from us, and successfully aligns AI systems that *they* create, I'd count those as "aligned" AIs.

("Aligned AIs" themselves will, of course, have the same values as evolved aliens. Benefiting their values would be the same as benefitting the values of some evolved aliens, so they don't need a separate category.)

Now, why do I separately consider the values of evolved aliens and misaligned AIs? There are two reasons.

Firstly, compared to AI, evolved aliens probably have minds that are more similar to ours, and face decisions that are more similar to ours. Thus, there's a stronger case that our decisions correlate more with their decisions, making the case for ECL-based cooperation stronger.

Second, AI progress is currently fast, and I have less than perfect confidence in humanity's ability to only create and empower aligned AI systems. This (ominously) suggests that we may soon have unique opportunities to benefit or harm the values of misaligned AI systems.