Should we expect moral convergence?

Plan for this article:

- Share with Pablo, Aron + wait a week or two
- Revise
- Share + discuss with GPI folk
- Revise
- Share with Open Phil, FHI, etc

Summary of most important responses from people:

- Various details of how to state the key claims
- " if convergence is very unlikely, you might take this to indicate that our own moral views are very unlikely to be on the right track, and that we are therefore not in a good position to tell what it would mean for the future to go well. Although presumably we could appeal to moral uncertainty here and replace "goes well" with "goes well in expectation," and your point would still hold."
- Would someone who believes in the idealised preference view accept that If the idealised preference view is true, you should not expect moral convergence.
    - I take it that if the view is roughly that what it is for something to be valuable is for agents to value it under suitably idealised conditions, then if convergence fails, what we get is a form of relativism (with value being relative to the agent). And I would have thought that, at least at the outset, those who endorse the idealisation view would recognise the avoidance of relativism as a desideratum.
    - But having had a very quick look at the literature, I'm no longer as confident of this:
    - - David Lewis (in "Dispositional Theories of Value") recognises the desideratum, but leaves it open whether it can be met.
    - - Michael Smith (in The Moral Problem) writes that "convergence in the hypothetical desires of fully rational creatures is required for the truth of normative reason claims" (p. 173) and that we have reason to "have some confidence...[that] there will be a convergence in our desires under conditions of full rationality" (p. 187).
    - - On the other hand, David Sobel (in chapter 5 of From Valuing to Value) denies that convergence is likely, but still endorses a subjectivist idealisation view
- Maybe worth including among "arguments for": Something like Railton-style naturalistic realism, where the moral truths are just the set of practices we non-randomly converge on. (I think my memory of Railton's view is badly oversimplified, but I remember finding the basic idea mildly plausible. But its plausibility depends in part on conditions that may not hold in the very-long-run future, e.g. rough parity of physical and intellectual capabilities among most agents.)
    - (A more general argument for convergence, that Railton's view exemplifies, is extrapolation from historical evidence, to the extent that you think there's been moral convergence and that this moral convergence is a result of shared but independent factors, rather than influence between societies/civilizations.)
    - Other metaethical views might also strongly predict convergence, e.g. Hobbes/Gauthier-style contractarianism.
    - I would classify both Railton and Hobbes/Gauthier as very far from full-blooded realism, and so it seems to me it's less important if convergence is true because one of these metaethical views is true, than if it's true because some more robustly realist view is true. But as you say, this attitude won't be widely shared.
    - *So the idea is: Take what we would, as a matter of fact, converge on. Rigidify that. That's what morality is. But then we get something weird when we can affect what we*

can converge on. It means that if we do A, then A is right and B wrong; and vice-versa. Also just very implausible as a view.

Warning: Very early thoughts. Speculative. Talks a bunch about optimising the universe, but if that's too crazy-seeming, that discussion could be replaced with the idea of building a just society for a very long time (e.g. until the Earth is no longer habitable) — many but not all of the considerations would port over.

Terminology

Let's distinguish between several hypotheses (kept vague for now):

- *Convergence hypothesis*: A wide variety of sufficiently good civilisations, over a sufficiently long time period, will end up having roughly the same preferences and beliefs.

If this is to be true in the way we want, both of the following need to be true:

- *Scientific convergence hypothesis*: A wide variety of sufficiently good civilisations, over a sufficiently long time period, will end up having roughly the correct empirical beliefs.
- *Moral convergence hypothesis*: A wide variety of sufficiently good civilisations, over a sufficiently long time period, will end up having roughly the correct preferences.

There are many bad ways in which this could be true, which I'll bundle all together into the following:

- *Power convergence hypothesis*: A wide variety of sufficiently good civilisations, over a sufficiently long time period, will end up having preferences and beliefs that are optimised for something roughly orthogonal to what would be optimised for under the correct preferences and beliefs. (This could be 'economic power', 'technological power', 'population size', etc.)

There are some different ways in which the Convergence hypothesis could be false:

- *Weak path-dependence hypothesis*: For a wide variety of civilisations, the preferences or beliefs that the civilisation ends up adopting is significantly dependent on the conditions of the early development of the civilisation.
- *Strong path-dependence hypothesis*: For a wide variety of sufficiently good civilisations, the preferences or beliefs that the civilisation ends up adopting is significantly depending on the conditions of the early development of the civilisation.
- *Weak random walk hypothesis*: For a wide variety of civilisations, the preferences or beliefs that the civilisation ends up adopting is basically random.
- *Strong random walk hypothesis*: For a wide variety of sufficiently good civilisations, the preferences or beliefs that the civilisation ends up adopting is basically random.

As you can see, there's a lot that needs to be made precise: how wide is 'wide', what counts as 'sufficiently' (in both instances), what counts as having 'roughly the same' preference and beliefs, what it means to ‘depend upon’ earlier conditions. Ultimately we’ll want to make these precise.

Note that the path dependence or random walk hypotheses are true if they are true of either beliefs or preferences. I take it that most people would think they are much more likely to be true of preferences, so I’ll focus on that in what follows. But bear in mind that there’s at least a large number of people (Kuhnians, critical theorists, scientific anti-realists) who might deny that, too.

I’m using ‘correct preferences’ broadly — plug in your favourite metaethical term instead if you want (e.g. “true moral beliefs”, “preferences I’d want, upon ideal reflection, for future civilisation to have”, “woo yeah preferences of type X” etc). I mean for all meta-ethical views other than nihilism to be on board with this idea.

By ‘depend upon’, I mean that they are to some extent predictable. If the final preferences of civilisation are determined by the starting preferences of civilisation, but it’s a chaotic system and you the starting preferences would give us no predictive power over the final preference, then this doesn’t count as ‘depending upon’ as I’m using the term.

By ‘sufficiently good’ civilisation I mean one that has avoided very clear lock-in scenarios over the next thousand years: no extinction, no unrecuperable civilisational collapse, no AI dictatorship.

Importance

My focus in this document is going to be on the moral convergence hypothesis. It’s important whether this is correct, for a couple of reasons:

First:

If it’s the case that moral convergence is highly probable (e.g. 50% likely), then it seems hard to deny that the best course of action is to ensure that we get civilisation to a good enough state (no extinction, no permanent civilisational collapse, no lock-in of values). Then we let future people take it from there.

If it’s the case that moral convergence is very unlikely (e.g. 1 in a billion, conditional on continued survival), then it becomes much more plausible that the best course of action is to ensure that the future goes well conditional on survival. (Mitigating extinction risk by 1 percentage point is only increasing the chance of achieving the best outcome by 1 in 10^-11. It seems plausible that you could do better than this by trying to influence how well the future goes, conditional on survival — for example by promoting a certain set of values.)

Here’s a particular comparison to make this vivid:

- Plan A (The default plan): Consume a most EA resources in reducing extinction risk over the next hundred years, accepting that we probably have minor influence over how civilisation goes in the long run.

- Plan B (The long game): Continually reinvest EA resources in whatever will have the highest long-term growth. Wait until we control a significant proportion of the world's resources. Then steer humanity in *exactly* the right direction.

The idea would be that, even if Plan B means that we miss 99% of opportunities to reduce existential risk, it's still the better than plan A because being able to influence how the future goes conditional on survival is so much better than 'merely' reducing existential risk.

Second:

If humanity goes extinct but higher mammals (or even other vertebrates) remain (e.g. because of a pandemic), then there is some significant chance that higher intelligence will evolve again in the next billion years of habitable life. If we expect moral convergence, then, depending on how well we think our civilisation is going (i.e. if we think it's going worse than one would have expected from a random draw of possible animal preferences), we might think that this new higher intelligence has as good or better a chance of getting to the moral truth.

Even if the entire planet is destroyed or rendered uninhabitable, there is some chance that aliens will spread to significant parts of the universe that is accessible-to-us. Again, if we expect moral convergence, then, depending on how well we think our civilisation is going (i.e. if we think it's going worse than one would have expected from a random draw of possible alien preferences)), we might think that this new higher intelligence has as good or better a chance of getting to the moral truth.

Third:

If humanity is replaced by artificial agents, which have general-purpose reasoning abilities, then if we think that there will be moral convergence then we should expect those artificial agents to converge on the moral truth.

Note:

If we think that a random set of preferences are, in expectation, as likely to produce a good outcome of given magnitude as they are to produce a bad outcome of that magnitude, we can effectively ignore the random walk hypotheses. However, many moral views won't have this sort of strong good/bad symmetry: They might think there are more types of ways the world could be bad than ways is could be good (or vice-versa), or that bad count for more, morally, than goods (or vice-versa). In which case the random walk hypotheses are relevant to our decision-making.

How plausible is moral convergence? - Arguments against

The disjunctive syllogism metaethical argument

I'll start off just by considering (i) internalist moral realism (ii) externalist moral realism and (iii) the idealised preference view.

Here's one argument:

- If the idealised preference view is true, you should not expect moral convergence.
    - What preferences you end up with are highly depending on features of what your particular brain is like.
    - Once we have extremely advanced technology, a universe optimised to satisfy one set of preferences will be very different from a universe optimised to satisfy even a slightly different set of preferences. Consider how different the universe would be if it were optimised for by the following theories, which are extremely close relative to the whole space of possible moral views:
        - Hedonistic vs preference utilitarianism
        - Biological-humans-only utilitarianism vs digital-minds-inclusive utilitarianism
        - (Potentially) Critical level vs total utilitarianism
            - (If it's a mild cost to unify experiences such that they form persons — rather than every experience moment being its own person — then the universe optimised for total utilitarianism would be astronomically bad according to critical level utilitarianism).
- If the internalist moral realist view is true, you should not expect moral convergence.
    - The final preferences we have will be very significantly influenced by the fact that it was monkeys who developed higher intelligence, rather than bees or lobsters or tigers, and the differences in preferences between humans, bees, lobsters and tigers is small compared to the space of all possible preference-structures that one could have.
    - It would seem fortuitous indeed if the preferences of homo sapiens just happen to be close enough to the correct preferences that we morally converge, even though other lifeforms wouldn't.
- If the externalist moral realist view is true, you should not expect moral convergence.
    - Because you've accepted there's no motivating force towards the true moral view.
- Therefore, you should not expect moral convergence.

Response:

I can see some ways in which the above arguments could be wrong:

- If the attraction for rational agents to believe the moral truth is *very strong* (e.g. as strong as mathematics), then maybe we'd end up in the same place wherever we started from.

    - On my favourite metaethics and normative ethics, there is a strong attractor: once you experience certain states, you see that they're good, and as long as you're motivated to do the right thing, you'll see that having more good stuff is the right thing to do. So as long as future people are generally morally motivated, and expose themselves to a wide variety of experiences, they'll converge on the right view.
        - But I'm almost certainly wrong about metaethics and normative ethics.
        - And, even so, if they're not *very* exploratory, they could still easily miss out almost all the value. Suppose that unusual state of matter X grounds experiences 100x better than any other.
        - And there could be other forces that push us away from morality.
- If there's some reason why the idealisation process pushes in the same direction for everyone.
    - Though I don't understand, on idealised preference metaethical views, how it could be that the idealisation process is objective, rather than a matter of one's preference over how idealisation should go.
        - Reason: If you're willing to have some universal objective standard for how an idealisation process should go, then you're committed to just the same spooky metaphysics as the more full-blooded moral realist, undercutting the reasons for having an ideal preference view in the first place.
    - And if what the idealisation process is *is* determined by the person's preferences over how the idealisation process should go, then I'd expect there to be radically different outcomes. Some people value constency, others don't (etc - Just look at moral philosophers!)

The base rate argument

Value fragility: The space of 'extremely good' ways of organising matter is astronomically small compared to the space of all ways of organising matter. (Like, probably way lower than 1 out of 100!, but the precise number doesn't really matter.)

If this is our base rate, then we need truly extraordinary evidence to think that a proportion of the universe will be turned into one of these best possible states.

Response:

- Updating from extremely low priors is confusing and it's not nearly as crazy as it seems to update from astronomically low base rates (consider: updating to the view that the sequence of dealt cards from a pack of card is thus-and-so involves moving from 1/52! to 90% or so).
- The argument proves too much. If God tells us that, yes, in the future the universe got optimised for the best thing — how surprised would you be? Not 1/100! surprised.)

How plausible is moral convergence? - Arguments for

Philosophical agnosticism

Philosophy is really difficult. So, no matter how many arguments we consider back and forth, we'll never have a very low credence (i.e. <1%) that the moral convergence hypothesis is false. And 1% credence is enough, for most people, to support extinction risk mitigation over making future better conditional on survival.

Response:

I have significant reservations about relying on credences where the only reason they are as high as they are is because it seems overconfident to go any lower. I'd have (much) more sympathy for the view that we have no idea whether moral convergence is true, so we should pursue a portfolio approach: significant investment in extinction risk reduction, and significant investment in improving the future conditional on survival.

Realism is what matters

Things only 'really' matter if full-blooded realism (i.e. internalist moral realism) is true. So we should act on the assumption that it is true.

Two ways of fleshing out things 'really' mattering: (i) all views other than internalist moral realism are really forms of error theory, and there is no moral value on those views; (ii) the intertheoretic comparison between (utiliarianism | internalist moral realism) and (utiliarianism | not-internalist moral realism) is not 1:1, instead it's thousands or millions to one.

Response:

I have real sympathy for this argument, but I expect it not to be widely appealing. Giving justification for the core claim would involve some serious philosophical work.

Literature that discusses this issue or related issues

- Beckstead - on convergence
- Christiano - on meeting aliens
- Shulman - morality not much of a sacrifice
- Yudkowsky - Value is fragile