Note to readers: This is an unpolished document of notes and references, put together by Michael Aird. It's associated with Convergence Analysis's "crucial questions for longtermists" project. Feel free to make comments or suggestions.

(Asterisks mark works that I - Michael - haven't properly read yet.)

Questions that are maybe worth including somewhere

One option would be to have a section for "Questions that are less important, less common, or less explored by me". Maybe this would list the questions, or maybe it would just be a link to a post or doc that does so.

Some additional potential topics can be found here:
https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities
And here:
https://forum.effectivealtruism.org/posts/6x2MjPXhpPpnatJFQ/some-promising-career-ideas-beyond-80-000-hours-priority
Though I've already integrated a decent portion of them into these docs.

Natural risks

Ord discusses these risks, and also has relevant questions in Appendix F which I haven't integrated here yet.

What is the total natural existential risk?

Could mention stuff about the argument for an "upper bound" to natural extinction risk based on how long humanity has survived so far, or how long mammals/species typically last before extinction, etc.

https://forum.effectivealtruism.org/posts/BAhWbAEmGieXvMHJd/pangea-the-worst-of-times

This is definitely important to a lot of longtermists' views. But is it something they disagree on? Or do they basically just all share the same view?

Bostrom writes:

> It may not be surprising that existential risks created by modern civilization get the lion's share of the probability. After all, we are now doing some things that have never been done on Earth before, and we are developing capacities to do many more such things. If non-anthropogenic factors have failed to annihilate the human species for hundreds of thousands of years, it could seem unlikely that such factors will strike us down in the next century or two. By contrast, we have no reason whatever not to think that the products of advanced civilization will be our bane.
>
> We shouldn't be too quick to dismiss the existential risks that aren't human-generated as insignificant, however. It' s true that our species has survived for a long time in spite of whatever such risks are present. But there may be an observation selection effect in play here. The question to ask is, on the theory that natural disasters sterilize Earth-like planets with a high frequency, what should we expect to observe? Clearly not that we are living on a sterilized planet. But maybe that we should be more primitive humans than we are? In order to answer this question, we need a solution to the problem of the reference class in observer selection theory [76]. Yet that is a part of the methodology that doesn' t yet exist. So at the moment we can state that the most serious existential risks are generated by advanced human civilization, but we base this assertion on direct considerations. Whether there is additional support for it based on indirect considerations is an open question

What are the pros and cons of efforts to reduce natural existential risks in general?

What I have in mind is that idea that, since we have reason to believe natural *extinction* risks must be fairly low, any interventions we do to reduce them might have a substantial chance of increasing risks.

E.g., human-caused deflection of asteroids. Ord discusses this. It's also discussed here and here.

[supervolcanoes]

https://forum.effectivealtruism.org/posts/BAhWbAEmGieXvMHJd/pangea-the-worst-of-times

[asteroids and comets]

[stellar explosions]

Less commonly discussed, or probably less important, specific risks or technologies

[“Other environmental damage” (not climate change)]

Discussed by Ord. He has some relevant questions in Appendix F, which I haven’t yet integrated here.

Benefits, risks, and best approaches to human enhancement/genetic engineering

MacAskill: “A related, but more general, argument, is that the most pivotal point in time is when we develop techniques for engineering the motivations and values of the subsequent generation (such as through AI, but also perhaps through other technology, such as genetic engineering or advanced brainwashing technology), and that we’re close to that point. (H/T Carl Shulman for stating this more general view to me).”

Turchin lists as an “extinction risk”: “Disjunction: Human enhancement leads to competing species”

https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors#Genetic_enhancement

https://forum.effectivealtruism.org/posts/T8eKL6xdfL4yA2kvg/genetic-enhancement-as-a-cause-area *

https://www.nickbostrom.com/papers/embryo.pdf *

https://www.gwern.net/Embryo-selection *

Risks related to evolutionary dynamics

https://www.nickbostrom.com/fut/evolution.html

Hanson has discussed this sort of thing

This looks relevant: https://reducing-suffering.org/the-future-of-darwinism/*

Cyber weapons / cybersecurity / information security stuff

https://forum.effectivealtruism.org/posts/6x2MjPXhpPpnatJFQ/some-promising-career-ideas-beyond-80-000-hours-priority#Information_security
And the resources linked to there.

There's a section on "Difficulty of destructive cyber attacks" in the "strategy variables" document.

Risks from extraterrestrial intelligence

Alexey Turchin, Global Catastrophic Risks Connected with Extra-Terrestrial Intelligence *

How likely is it that our observable universe contains extraterrestrial intelligence (ETI)?

See “How likely is it that our observable universe contains extraterrestrial intelligence (ETI)? How valuable would a future influenced by them rather than us be?” elsewhere.

If there are ETI in our observable universe, do they pose risks to us? If so, how large, and via what pathways?

Could METI/“passive SET” pose risks?

Ord discusses this

Turchin makes I think a similar point, listing under “remote and hypothetical [extinction] risks” “Downloading Alien AI via SETI”, and also “METI attracts dangerous ET”.

What can we do to reduce risks from ETIs? How tractable are these options?

Whole brain emulations

Perhaps this should be included as part of the AI cluster of questions?

Age of Em

https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors

https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities#Whole_brain_emulation

https://web.archive.org/web/20191224210426/https://www.researchgate.net/publication/285980771_Is_Brain_Emulation_Dangerous *

The Prospects of Whole Brain Emulation within the next Half- Century

There's a relevant section in the "strategy variables" document.

Denis Drescher suggested adding the following questions about WBE:

- Will ems arrive before AGI? If not, will there be any incentive to still develop ems?
- Will ems supersede biological humans or exist in parallel with them?
- Will ems speed the development of AGI?
- Will ems affect geopolitical in-/stability?
- Will two transitions (status quo -> ems -> AGI) be more or less risky than one transition (status quo -> AGI)
- Will ems cause a greater or lesser expected degree of settlement of the affectable universe? (Less incentive, greater aptitude, potentially short interlude before AGI.)
- How positive or negative will em lives be?
- How will wild animals coexist with em cities or on dyson spheres?

(I think Robin Hanson thinks that ems will be first and Brian Tomasik thinks AGI will be first.)

Malevolent humans

https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors

https://www.lesswrong.com/posts/Ft2Cm9tWtcLNFLrMw/notes-on-the-psychology-of-power

https://www.lesswrong.com/posts/ktr39MFWpTqmzuKxQ/notes-on-psychopathy

I also have some other unpolished notes on this.

Intelligence Amplification

There's a relevant section in the "strategy variables" document.

Wisdom Amplification

There's a relevant section in the "strategy variables" document.

Brain computer interfaces

A New X-Risk Factor: Brain-Computer Interfaces - EA Forum

[Stuff about space governance]

Space governance is important, tractable and neglected

Off-Earth Governance *

https://forum.effectivealtruism.org/tag/space

[Quantum computing stuff?]

https://hearthisidea.com/episodes/jaime/

[Lethal autonomous weapons, drone swarms, etc.?]

I feel like this isn't *super* important, and the ways it's important are probably mostly best captured in the AI section. But I'm not certain.

Why those who care about catastrophic and existential risk should care about autonomous weapons

[Pros and cons of more government power/centralisation/surveillance]

This is definitely very important, but maybe it's best integrated into somewhere else (or multiple places), rather than as its own separate category.

This is more like a risk/security factory than like a risk itself.

There's a bunch of relevant stuff in "Importance of, and best approaches to, totalitarianism and dystopias". The idea with this separate question/topic would be to also consider the ways that this increased power/centralisation/surveillance could help *reduce* other risks.

Maybe that's better considered as part of "Are there downsides to pursuing existential security? What are they? How large are they?"

https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities#Global_governance

https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities#Surveillance

https://www.effectivealtruism.org/articles/ea-global-2018-the-future-of-surveillance/

https://www.ted.com/talks/nick_bostrom_how_civilization_could_destroy_itself_and_4_ways_we_could_prevent_it/transcript

[Recommender systems at top tech firms]

https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities#Recommender_systems_at_top_tech_firms

https://forum.effectivealtruism.org/posts/xzjQvqDYahigHcwgQ/aligning-recommender-systems-as-cause-area-1

[Making investing for the long-term easier/more effective]

https://forum.effectivealtruism.org/posts/xoxbDsKGvHpkGfw9R/problem-areas-beyond-80-000-hours-current-priorities#We_may_need_to_invest_more_to_tackle_future_problems

Simulation stuff?

Does this affect anyone's decisions?

https://www.lesswrong.com/posts/Qz6w4GYZpgeDp6ATB/beyond-astronomical-waste

Turchin (in one of his maps) gives as an "improbable idea":

> "Control of the simulation (if we are in it)
>
> - Live an interesting life so our simulation isn't switched off
> - Don't let them know that we know we live in simulation
> - Hack the simulation and control it
> - Negotiation with the simulators or pray for help"

Turchin lists under "remote and hypothetical risks" " We live in simulation and it switched off or is built to model catastrophes"

MacAskill writes:

> "(2) The case for focusing on AI safety and existential risk reduction is much weaker if you live in a simulation than if you don't. (In general, I'd aver that we have very little understanding of the best things to do if we're in a simulation, though there's a lot more to be said here.)
>
> The primary reasons for believing (2) are that if we're in a simulation it's much more likely that the future is short, and that extending our future doesn't change the total amount of lived experiences (because the simulators will just run some other simulation afterwards), and that we're missing some crucial consideration around how to act."

He links to this:
https://longtermrisk.org/how-the-simulation-argument-dampens-future-fanaticism#Simulation_argument_upshifts_the_relative_importance_of_short-term_helping

Schulman replies:

> “I would note that the creation of numerous simulations of HoH-type periods doesn't reduce the total impact of the actual HoH folk. E.g. say that HoH folk get to influence 10^60 future people, and also get their lives simulated 10^50 times (with no ability to impact things beyond their own lives), while folk in a non-HOH Earthly period get to influence 10^55 future people and get simulated 10^42 times. Because simulations account for a small minority of the total influence, the expected value of an action (or the evidential value of a strategy across all like minds) is still driven primarily by the non-simulated cases. Seeming HoH folk may be simulated more often, but still have most of their influence through unsimulated shaping of history.
>
> If simulations were so numerous that most of the value in history lay in simulations, rather than in basement-level influence, then things might be different. But I think argument #3 doesn't work for this reason.”

https://www.nickbostrom.com/existential/risks.html:

> A case can be made that the hypothesis that we are living in a computer simulation should be given a significant probability [27]. The basic idea behind this so-called “ Simulation argument” is that vast amounts of computing power may become available in the future (see e.g. [28,29]), and that it could be used, among other things, to run large numbers of fine-grained simulations of past human civilizations. Under some not-too-implausible assumptions, the result can be that almost all minds like ours are simulated minds, and that we should therefore assign a significant probability to being such computer-emulated minds rather than the (subjectively indistinguishable) minds of originally evolved creatures. And if we are, we suffer the risk that the simulation may be shut down at any time. A decision to terminate our simulation may be prompted by our actions or by exogenous factors.
>
> While to some it may seem frivolous to list such a radical or “philosophical” hypothesis next the concrete threat of nuclear holocaust, we must seek to base these evaluations on reasons rather than untutored intuition. Until a refutation appears of the argument presented in [27], it would intellectually dishonest to neglect to mention simulation-shutdown as a potential extinction mode

He also has other relevant parts in that paper.

Anthropics stuff

Does this affect anyone’s decisions?

Dissolving fermi paradox is somewhat relevant, and this is somewhat relevant to aliens. This could maybe be included as a subquestion of an aliens question?

This also seems relevant to things like estimating x risks and what we can learn from past GCRs or near-misses.

Beard et al.:

> "Another problem with these methods is that some events are excluded from the historical record because of the 'anthropic shadow' they would leave. Roughly speaking, were any such event to have occurred in the past, this would have led to the non-existence of the present observer. Therefore, even if its probability was extremely high, it must seem to us as if it could not have happened because our very existence depends on it (Ćirković, Sandberg, & Bostrom , 2010). Manheim (2018) looked at risk estimates of natural pandemics and concluded that there is significant uncertainty about the relationship between historical patterns and present risk because of such "anthropic factors and other observational selection biases". Tegmark and Bostrom (2005 – source 65) also take account of this effect when quantifying the threat from particle physics experiments, but most theories ignore it."

https://www.nickbostrom.com/existential/risks.html has a section on Observation selection effects.

Bostrom also writes:

> It may not be surprising that existential risks created by modern civilization get the lion's share of the probability. After all, we are now doing some things that have never been done on Earth before, and we are developing capacities to do many more such things. If non-anthropogenic factors have failed to annihilate the human species for hundreds of thousands of years, it could seem unlikely that such factors will strike us down in the next century or two. By contrast, we have no reason whatever not to think that the products of advanced civilization will be our bane.
>
> We shouldn't be too quick to dismiss the existential risks that aren't human-generated as insignificant, however. It' s true that our species has survived for a long time in spite of whatever such risks are present. But there may be an observation selection effect in play here. The question to ask is, on the theory that natural disasters sterilize Earth-like planets with a high frequency, what should we expect to observe? Clearly not that we

are living on a sterilized planet. But maybe that we should be more primitive humans than we are? In order to answer this question, we need a solution to the problem of the reference class in observer selection theory [76]. Yet that is a part of the methodology that doesn' t yet exist. So at the moment we can state that the most serious existential risks are generated by advanced human civilization, but we base this assertion on direct considerations. Whether there is additional support for it based on indirect considerations is an open question

https://arxiv.org/pdf/astro-ph/0512204.pdf*

“Normative”/ethics stuff

How should we handle “Pascal’s muggings” and/or potential “fanaticism”?

Not sure this really counts as normative/ethical; might be more like a rationality thing.

Does this influence different decisions *among longtermists* (rather than differences between lngntermists and others)? I can imagine it playing a role in things like narrow (e.g. MIRI) vs broad (e.g. improving institutional decision-making) approaches, perhaps?

GPI research agenda:

> Mitigation of catastrophic risk is sometimes a matter of an extraordinarily small chance of generating extraordinarily high value. Is expected utility theory the correct approach for dealing with decisions of this character (Bostrom 2009; Tarsney 2018) (INFORMAL: Yudkowsky 2013)? Does any plausible alternative lead away from the idea that the opportunities in question are among the best from an ex ante evaluative standpoint (INFORMAL: Karnofsky 2011)?
>
> [...]
>
> Under moral uncertainty, do some axiological views with very high stakes swamp the expected value calculation? If so, which views are they? What is the best way to deal with this ‘fanaticism’ issue? (Ross 2006; MacAskill and Ord 2018; Cotton-Barratt and Greaves MS) (INFORMAL: MacAskill 2018-a)

[Infinite ethics]

Not sure if this actually does influence differing views. But it definitely *could*.

Discussed a bit here.

I believe Nick Bostrom and Amanda Askell have discussed some relevant points?

GPI research agenda:

> Let finitism be the claim that, even if we ought perhaps to aim to bring about an astronomically large finite amount of value in the future, we ought not to aim explicitly to bring about an infinitely large amount of value. Is finitism defensible? If it is not defensible, is this a reductio of the idea that we ought to try to bring about an astronomically large finite amount of value, or an argument that we really should be pursuing infinite amounts of value? If the latter, how do we compare outcomes involving possibilities of infinite quantities of value, in order to decide which such outcomes to pursue? (Vallentyne and Kagan 1997; Basu and Mitra 2003; Vallentyne and Lauwers 2004; Zame 2007; Asheim 2010; Bostrom 2011; Arntzenius 2014) (INFORMAL: West 2015)

[Population ethics?]

The fact that we're taking longtermism as a starting point makes this much less important. But there still might be differences in how far towards total utilitarianism rather than e.g. person-affecting people are, even if the vast majority at least lean in roughly the same direction.

More notably, this affects different priorities for negative-leaning/suffering-focused people.

GPI research agenda:

> Should we be more concerned about avoiding the worst possible outcomes for the future than we are for ensuring the very best outcomes occur (whether because the worst outcomes are worse than the best outcomes are good, because avoidance of the bad outcomes is more neglected, or because bad outcomes should be weighted more than good outcomes when other relevant things are equal) (Hurka 2010) (INFORMAL: Althaus and Gloor 2018; Tomasik 2018)? If so, what activities would be best? (MacAskill MS-a) (INFORMAL: Gloor 2018)
>
> [...]
>
> What do the most plausible person-affecting views in population ethics say about the value of reducing extinction risk? (INFORMAL: Greaves 2016)

https://longtermrisk.org/descriptive-population-ethics-and-its-relevance-for-cause-prioritization/
https://forum.effectivealtruism.org/posts/CmNBmSf6xtMyYhvcs/descriptive-population-ethics-and-its-relevance-for-cause

[Discounting?]

Maybe not worth mentioning as longtermists pretty much be definition don't do substantially pure time discounting? But I guess some might discount not at all, while others do so at least slightly (so they e.g. care about thousands or millions of years, but not more)?

Moral status of non-humans???

Animals, AIs, aliens, future evolved life, etc.

I now think it probably makes sense to just have this be part of things like "How close to the appropriate size are influential agents' moral circles likely to be?"

About AI

These are additional questions/topics that aren't currently in my list of AI-related questions in the main docs.

How likely are AI race dynamics? How dangerous would those be? How can they be mitigated?

This is definitely important, but it's also possible we don't want to draw attention to these questions because doing so could itself be an attention hazard.

If we do include questions like this, it's possible they should be split up.

GovAI Research Agenda:

> Even the mere perception by governments and publics of such military (or economic) potential could lead to a radical break from the current technology and world order: shifting AI leadership to governments, giving rise to a massively funded AI race and potentially the securitization of AI development and capabilities. This could undermine the liberal world economic order. The intensity from a race dynamic could lead to catastrophic corner-cutting in the hurried development and deployment of (unsafe) advanced AI systems. This danger poses extreme urgency, and opportunity, for global cooperation.

["specialized machines always beat general ones"]

Yann LeCunn responds to the instrumental convergence argument with, among other things:

> "A second machine, designed solely to neutralize an evil super-intelligent machine will win every time, if given similar amounts of computing resources (because specialized machines always beat general ones)."

[Intelligence will naturally involve a desire to dominate?]

Yann LeCunn counters this view:

> “We dramatically overestimate the threat of an accidental AI takeover, because we tend to conflate intelligence with the drive to achieve dominance. [...] But intelligence per se does not generate the drive for domination, any more than horns do.”

But I haven’t heard any AI safety researchers actually express the view LeCunn critiques

[Hatred, malice, desire for vengeance, or similar would be required for an AI to do catastrophically bad things]

Stuart Russell counters this view:

> “It is trivial to construct a toy MDP in which the agent's only reward comes from fetching the coffee. If, in that MDP, there is another "human" who has some probability, however small, of switching the agent off, and if the agent has available a button that switches off that human, the agent will necessarily press that button as part of the optimal solution for fetching the coffee. No hatred, no desire for power, no built-in emotions, no built-in survival instinct, nothing except the desire to fetch the coffee successfully.” And also: “The point is that the behaviors we are concerned about have nothing to do with putting in emotions of survival, power, domination, etc. So arguing that there's no need to put those emotions in is completely missing the point.”

I don’t think I’ve seen “serious” critics of AI risk arguments clearly holding this critiqued view. But it may be implicit sometimes. And definitely a large portion of journalists and the general public seem to hold this view.

Suggestive evidence that this view is implicitly held by serious researchers: LeCunn writes

> “A second machine, designed solely to neutralize an **evil** super-intelligent machine will win every time, if given similar amounts of computing resources (because specialized machines always beat general ones)”. And “Bottom line: there are lots and lots of ways to protect against badly-designed intelligent machines turned **evil**.” (Emphases added.)

Misc

Best indices/proxies for tracking how near-term effects might influence longterm outcomes?

GPI research agenda:

> 1.6 Economic indices for longtermists
>
> It is standard practice in economics to evaluate policies, explicitly or implicitly, on the basis of their expected short-term impact on total economic output. It is also standard, though less common, to evaluate policies on the basis of their expected short-term impact on economic indices designed to correspond more closely with human welfare, such as the human development index (HDI). From a longtermist perspective, however, the true measure of a policy's success is its impact on the long-term prospects of human civilisation. We must therefore ask how well the former indices track the latter objective, and, perhaps, how to construct and implement economic indices that track the latter objective more closely.
>
> Potential research projects:
>
> Much government policy, economic research, and philanthropic activity is intended ultimately to increase the general rate of economic growth. Economic growth could be extremely beneficial, from a long-term perspective, as it promises to improve the entire course of the future. However technology-driven growth may raise existential risks, due for example to nuclear accidents, engineered pandemics or artificial superintelligence (INFORMAL: Yudkowsky 2013), and growth in general may have other negative effects (for instance, risks to human life (Jones 2016), climate change (IPCC 2014), or meat consumption (INFORMAL: Bogosian 2015)). How radically do these drawbacks render growth an imperfect proxy for expected long-term wellbeing? Is the correlation between consumption growth and long-term wellbeing even positive, given the current drivers of growth, from a geographical, sectoral and technological perspective? (Friedman 2006; Cowen 2007; Tomasik 2013; Cowen 2018) (INFORMAL: Beckstead 2014)
>
> Of the comprehensive macroeconomic indices already available to us, which serve best as proxies for long-term expected global welfare (including but not limited to considerations of existential risks)? What would be the broad policy implications of targeting such indices instead of GDP per capita?
>
> Are there any promising proxies for long-term wellbeing not already tracked as macroeconomic indices (INFORMAL: Shulman 2013; Bostrom 2014)? If so, how could these proxies be formalised and measured, and what would be the broad policy implications of targeting them instead of GDP per capita?
>
> [...]

1.8 Longtermist status of interventions that score highly on short-term metrics

It is sometimes argued that interventions designed to score highly on short-term metrics—such as cost-effective poverty alleviation programmes—are also typically the actions with the best expected long-term consequences. If that is correct, then longtermism (even if true) has little practical significance. It is therefore important to evaluate this argument.

Potential research projects:

Is there any motivation for prioritising interventions that score highly on short-term metrics that is respectable from a longtermist perspective? (INFORMAL: Karnofsky 2014; Tomasik 2015)

To what extent should a worry of 'suspicious convergence' (INFORMAL: Lewis 2016) incline us against the hypothesis that the interventions that have the best short-termist motivation also fare well by longtermist lights?

What are the long-term effects of interventions that seem particularly high-priority from a short-term perspective, such as saving human lives (INFORMAL: Karnofsky 2013) or improving the conditions of caged hens (Matheny and Chan 2005) (INFORMAL: Shulman 2013)? What is the sign of these effects, and how substantial are they? Under what conditions, if any, might they exceed the expected long-term impacts of (other) efforts aimed explicitly at improving the long term?

https://reflectivedisequilibrium.blogspot.com/2013/12/what-proxies-to-use-for-flow-through.html

Something cross-cutting about differential progress, how advancing beneficial tech might advance risky tech, etc.?

But again, maybe this is best addressed in its specific instantiations, rather than as a cross-cutting thing.

If we do include something like this, here's my list of *some* sources worth linking to: https://forum.effectivealtruism.org/posts/EMKf4Gyee7BsY2RP8/michaela-s-shortform?commentId=xrj9XYjvsGLz6R2i6

The "strategy variables" document's section on "technological interdependency" is relevant.

Inside vs outside views, epistemic modesty, etc.

List of sources: https://www.lesswrong.com/posts/gcEayv6HtBogfov2n/michaela-s-shortform?commentId=nFrAa4zkCzhRgBMLH

GPI research agenda:

> 2.2 Epistemological issues
> Thinking about global prioritisation, particularly (although not only) within the longtermist paradigm, tends to rely on heavily philosophical considerations and to reach some surprising and counterintuitive conclusions. We must therefore assess the extent to which this unusual circumstance should undermine our confidence in the conclusions in question.
>
> Potential research projects:
>
> To what extent should an actor place weight on her own idiosyncratic 'inside view' judgments, rather than deferring to the views of the majority of peers/experts on the issue? (Elga 2007; Christensen 2007; Christensen 2009; Feldman and Warfield 2010; Wilson 2010; Christensen and Lackey 2013) (INFORMAL: Beckstead 2013; Lewis 2017)
>
> How much weight should we place on philosophical arguments? Is there a sound 'pessimistic induction' against placing much weight on them, assuming that most philosophical arguments in the past have been mistaken?
>
> What mechanisms can induce individuals to report their moral views honestly to each other?
>
> Should one have the same levels of epistemic modesty about unusual moral views as one should about unusual empirical views?

Something about the tractability of changing the course of history

Like what's discussed here

Would also be similar to some parts of "Weirdness of reality" from the "strategic variables" document, I think.

Importance of, and best approaches to, downside risks

But maybe this is best addressed in its specific forms, such as information hazards and the possibility of AI safety research enhancing capabilities, rather than as a cross-cutting thing?

If we do include something like this, here's my list of sources worth linking to: https://forum.effectivealtruism.org/posts/EMKf4Gyee7BsY2RP8/michaela-s-shortform?commentId=GLJdSeFpLQhg6p8cK

Even miscier

- Something about how much difference in impact between charities, approaches, etc. we should see as plausible (related to things like how efficient the charitable market is).
    - It seems like *maybe* this could inform disagreements between "MIRI-type people" and more "moderate" or "common sense" longtermists, the latter of which might agree AI seems more important but advocate for much more diversification because they see it as implausible that AI is *so so much* more important.
    - Maybe I should read the things linked to here: https://forum.effectivealtruism.org/posts/knJJvp5JGGSdy6ocr/assumptions-about-the-far-future-and-cause-priority?commentId=ECXXq6NHxFJjxXTi4

- Justin's variance idea
- Justin's ratio idea? Or should that be elsewhere?
- "Social group efficiency" from the "strategy variables" document
- "Degree of fragility of society" from the "strategy variables" document
    - I tentatively believe that this isn't really a key question aside from the ways it relates to other questions already mentioned in this series.
- "Intuition vs logic" from the "strategy variables" document
- "Offense vs defense" from the "strategy variables" document

Questions there's wide agreement upon

I could maybe make a section with a name along these lines which is for questions that *would* be crucial questions for longtermists, and whose answers *are* very important, but which the vast majority of longtermists seem to agree on. There may be major disagreement between longtermists and others on these questions, but not *among longtermists*.

Some candidates can be found scattered throughout the various docs related to this series. For example:

- What is the total natural existential risk? (more specifically, the idea that this must be fairly low, given that we've survived this long, typical mammalian species lifespan, etc.)

One tricky thing is that I think there are some questions (like the natural risks one) which some longtermists simply haven't noticed, or arguments they haven't heard. So no one really "disagrees", but there *are* people who don't really have an explicit view on the matter at all, and thus make different decisions (e.g., focusing on supervolcanoes). Not sure whether such questions should be in the main posts or in this section,

Relevant quote from Cottier & Shah:

> This does not decompose arguments exhaustively. It does not include every reason to favour or disfavour ideas. Rather, it is a set of key hypotheses and relationships with other hypotheses, problems, solutions, models, etc. Some examples of important but apparently uncontroversial premises within the AI safety community: orthogonality, complexity of value, Goodhart's Curse, AI being deployed in a catastrophe-sensitive context.

There are also relevant questions at the bottom of Notes regarding the Crucial questions for longtermists project.