Note to readers: This is a slightly outdated draft of an article I've now posted to the EA Forum: https://forum.effectivealtruism.org/posts/wicAtfihz2JmPRgez/crucial-questions-for-longtermists Feel free to comment on this doc with suggestions of additional questions, or suggestions of changes to the questions we have or how they're structured. I (Michael Aird) will continue to look at suggestions and potentially make edits to the EA Forum post.

This post was written for Convergence Analysis. It introduces a collection of "crucial questions for longtermists": important questions about the best strategies for improving the long-term future. This collection is intended to serve as an aide to thought and communication, a kind of research agenda, and a kind of structured reading list.

Introduction

The last decade saw substantial growth in the amount of attention, talent, and funding flowing towards existential risk reduction and longtermism. There are many different strategies, risks, organisations, etc. to which these resources could flow. How can we direct these resources in the best way? Why did they flow to the precise places they did? Are people able to effectively understand and critique the beliefs *underlying* various views - including their own - regarding how best to put longtermism into practice?

Relatedly, the last decade also saw substantial growth in the amount of research and thought on issues important to longtermist strategies. But this is scattered across a wide array of articles, blogs, books, podcasts, videos, etc. Additionally, these pieces of research and thought often use different terms for similar things, or don't clearly highlight how particular beliefs, arguments, and questions fit into various bigger pictures. This can make it harder to get up to speed with, form independent views on, and collaboratively sculpt the vast landscape of longtermist research and strategy.

To help address these issues, this post collects, organises, highlights connections between, and links to sources relevant to a large set of the "crucial questions" for longtermists.[1] These are questions whose answers *might* be "crucial considerations" - that is, considerations which are "likely to cause a major shift of our view of interventions or areas".

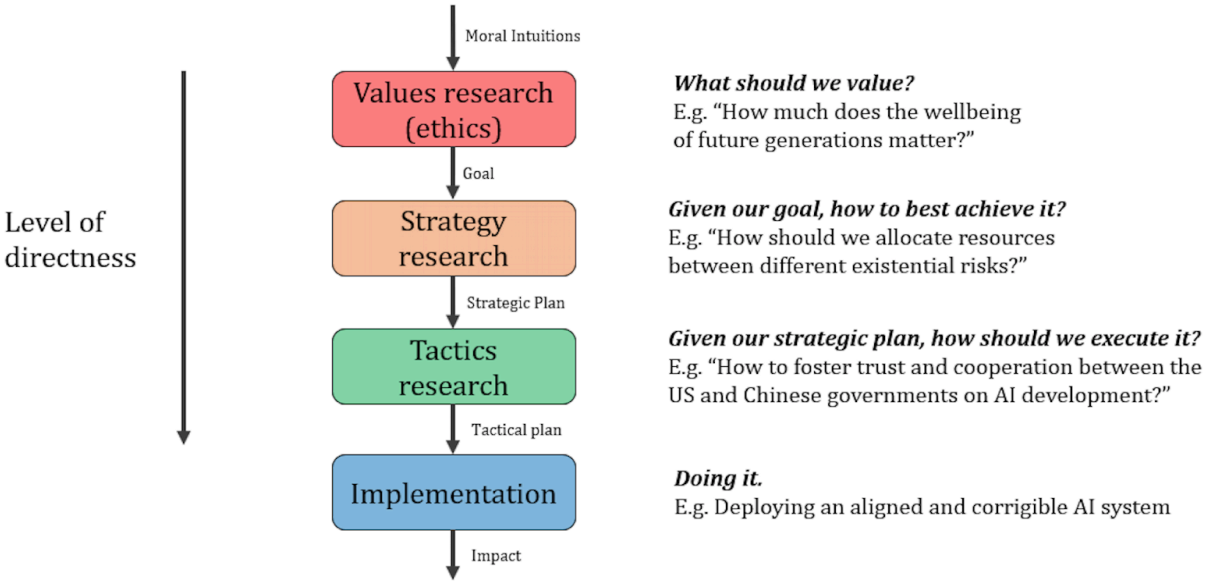
We collect these questions into topics, and then progressively then progressively break "top-level questions" down into the lower-level "sub-questions" that feed into them. For example, the topic "Optimal timing of work and donations" includes the top-level question "How will

[1] Most of the questions we cover are actually also relevant to people who are focused on existential risk reduction for reasons unrelated to longtermism (e.g., due to person-affecting arguments, and/or due to assigning sufficiently high credence to near-term technological transformation scenarios). However, for brevity, we will often just refer to "longtermists" or "longtermism".

‘leverage over the future” change over time?’, which is broken down into (among other things) “How will the neglectedness of longtermist causes change over time?” We also link to Google docs containing many relevant links and notes.

What kind of questions are we including?

The post A case for strategy research visualised the “research spine of effective altruism” as follows:



This post can be seen as collecting questions relevant to the “strategy” level.

One could imagine a version of this post that “zooms out” to discuss crucial questions on the “values” level, or questions about cause prioritisation as a whole. This might involve more emphasis on questions about, for example, population ethics, the moral status of nonhuman animals, and the effectiveness of currently available global health interventions. But here we instead (a) mostly set questions about morality aside, and (b) take longtermism as a starting assumption.[2]

One could also imagine a version of this post that “zooms in” on one specific topic we provide only a high-level view of, and that discusses that in more detail than we do. This could be

[2] Of course, some questions about morality *are* relevant even if longtermism is taken as a starting assumption. This includes questions about how important reducing suffering is relative to increasing happiness, and how much moral status various beings should get. Thus, we *will* touch on such questions, and link to some relevant sources. But we’ve decided to not include such questions as part of the core focus of this post.

considered to be work on “tactics”, or on “strategy” within some narrower domain. An example of something like that is the post Clarifying some key hypotheses in AI alignment. That sort of work is highly valuable, and we’ll provide many links to such work. But the scope of this post itself will be restricted to the relatively high-level questions, to keep the post manageable and avoid readers (or us) losing sight of the forest for the trees.[3]

Finally, we’re mostly focused on:

- Questions about which different longtermists have different beliefs, with those beliefs playing an explicit role in their strategic views and choices
- Questions about which some longtermists think learning more or changing their beliefs would change their strategic views and choices
- Questions which it appears some longtermists haven’t noticed at all, the noticing of which might influence those longtermists’ strategic views and choices

These can be seen as questions that reveal a “double crux” that explains the different strategies of different longtermists. We thus exclude questions about which practically, or by definition, all longtermists agree.

A high-level overview of the crucial questions for longtermists

Here we provide our current collection and structuring of crucial questions for longtermists. The linked Google docs contain some further information and a wide range of links to relevant sources, and I intend to continue adding new links in those docs for the foreseeable future.

“Big picture” questions (i.e., not about *specific* technologies, risks, or risk factors)

See here for notes and links related to these topics.

- Value of, and best approaches to, existential risk reduction
    - How “good” might the future be, if no existential catastrophe occurs?[4]
        - What is the possible scale of the human-influenced future?
        - What is the possible duration of the human-influenced future?
        - What is the possible quality of the human-influenced future?

[3] For example, we get as fine-grained as “How likely is counterforce vs. countervalue targeting [in a nuclear war]?”, but not as fine-grained as “Which precise cities will be targeted in a nuclear war?” We acknowledge that there’ll be some arbitrariness in our decisions about how fine-grained to be.

[4] Some of these questions are more relevant to people who haven’t (yet) accepted longtermism, rather than to longtermists. But all of these questions *can* be relevant to certain strategic decisions by longtermists. See the linked Google doc for further discussion.

            - How does the “difficulty” or “cost” of creating pleasure vs. pain compare?
        - Can and will we expand into space? In what ways, and to what extent? What are the implications?
            - Will we populate colonies with (some) nonhuman animals, e.g. through terraforming?
        - Can and will we create sentient digital beings? To what extent? What are the implications?
            - Would their experiences matter morally?
            - Will some be created accidentally?
    - How “bad” would the future be, if an existential catastrophe occurs? How does this differ between different existential catastrophes?
        - How likely is future evolution of moral agents or patients on Earth, conditional on (various different types of) existential catastrophe? How valuable would that future be?
        - How likely is it that our observable universe contains extraterrestrial intelligence (ETI)? How valuable would a future influenced by them rather than us be?
    - How high is total existential risk? How will the risk change over time?[5]
    - Where should we be on the “narrow vs. broad” spectrum of approaches to existential risk reduction?
    - To what extent will efforts focused on global catastrophic risks, or smaller risks, also help with existential risks?
- Value of, and best approaches to, improving aspects of the future other than whether an existential catastrophe occurs[6]
    - What probability distribution over various trajectories of the future should we expect?[7]
        - How good have trajectories been in the past?
        - How close to the appropriate size should we expect influential agents’ moral circles to be “by default”?
        - How much influence should we expect altruism to have on future trajectories “by default”?[8]

[5] See also our Database of existential risk estimates.
[6] This category of strategies for influencing the future could include work aimed towards shifting some probability mass from “ok” futures (which don’t involve existential catastrophes) to especially excellent futures, or shifting some probability mass from *especially awful* existential catastrophes to somewhat “less awful” existential catastrophes. We plan to discuss this category of strategies more in an upcoming post. We mean this to contrast with strategies aimed towards shifting probability mass from “some existential catastrophe” to “no existential catastrophe” (i.e., most existential risk reduction work).
[7] This includes things like how likely “ok” futures are relative to especially excellent futures, and how likely *especially awful* existential catastrophes are relative to somewhat “less awful” ones.
[8] This is about altruism in a general sense (i.e., concern for the wellbeing of others), not just EA specifically.

        - How likely is it that self-interest alone would lead to good trajectories “by default”?
    - How does speeding up development affect the expected value of the future?[9]
        - How does speeding up development affect existential risk?
        - How does speeding up development affect astronomical waste? How much should we care?
            - With each year that passes without us taking certain actions (e.g., beginning to colonise space), what amount or fraction of resources do we lose the ability to ever use?
            - How morally important is losing the ability to ever use that amount or fraction of resources?
        - How does speeding up development affect other aspects of our ultimate trajectory?
        - What are the best actions for speeding up development? How good are they?
    - Other than speeding up development, what are the best actions for improving aspects of the future other than whether an existential catastrophe occurs? How valuable are those actions?
        - How valuable are various types of moral advocacy? What are the best actions for that?
    - How “clueless” are we?
    - Should we find claims of convergence between effectiveness for near-term goals and effectiveness for improving aspects of the future other than whether an existential catastrophe occurs “suspicious”? If so, how suspicious?
- Value of, and best approaches to, work related to “other”, unnoticed, and/or unforeseen risks, interventions, causes, etc.
    - What are some plausibly important risks, interventions, causes, etc. that aren’t mentioned in the other “crucial questions”? How should the answer change our strategies (if at all)?
    - How likely is it that there are important unnoticed and/or unforeseen risks, interventions, causes, etc.? What should we do about that?
        - How often have we discovered new risks, interventions, causes, etc. in the past? How is that rate changing over time? What can be inferred from that?
        - How valuable is “horizon-scanning”? What are the best approaches to that?
- Optimal timing for work/donations
    - How will “leverage over the future” change over time?

[9] This refers to actions that speed development up in a general sense, or that “merely” change *when* things happen. This should be distinguished from changing *which* developments occur, or differentially advancing some developments relative to others.

        - What should be our prior regarding how leverage over the future will change? What does the “outside view” say?
        - How will our knowledge about what we should do change over time?
        - How will the neglectedness of longtermist causes change over time?
        - What “windows of opportunity” might there be? When might those windows open and close? How important are they?
        - Are we biased towards thinking the leverage over the future is currently unusually high? If so, how biased?
            - How often have people been wrong about such things in the past?
        - If leverage over the future *is* higher at a later time, would longtermists notice?
    - How effectively can we “punt to the future”?
        - What would be the long-term growth rate of financial investments?
        - What would be the long-term rate of expropriation of financial investments? How does this vary as investments grow larger?
        - What would be the long-term “growth rate” from other punting activities?
        - Would the people we’d be punting to act in ways we’d endorse?
    - Which “direct” actions might have compounding positive impacts?
    - Do marginal returns to “direct work” done within a given time period diminish? If so, how steeply?
- Tractability of, and best approaches to, estimating, forecasting, and investigating future developments
    - How good are people at forecasting future developments in general?
        - How good are people at forecasting impacts of technologies?
            - How often do people over- vs. underestimate risks from new tech? Should we think we might be doing that?
    - What are the best methods for forecasting future developments?
- Value of, and best approaches to, communication and movement-building
    - When should we be concerned about information hazards? How concerned? How should we respond?
    - When should we have other concerns or reasons for caution about communication? How should we respond?
    - What are the pros and cons of expanding longtermism-relevant movements in various ways?
        - What are the pros and cons of people who lack highly relevant skills being included in longtermism-relevant movements?
        - What are the pros and cons of people who don’t work full-time on relevant issues being included in longtermism-relevant movements?
- Comparative advantage of longtermists
    - How much impact should we expect longtermists to be able to have as a result of being more competent than non-longtermists? How does this vary between different areas, career paths, etc.?

        - Generally speaking, how competent, “sane”, “wise”, etc. are existing society, elites, “experts”, etc?
    - How much impact should we expect longtermists to be able to have as a result of having “better values/goals” than non-longtermists? How does this vary between different areas, career paths, etc.?
        - Generally speaking, how aligned with “good values/goals” (rather than with worse values, local incentives, etc.) are the actions of existing society, elites, “experts”, etc.?

Questions about emerging technologies

See here for notes and links related to these topics.

- Value of, and best approaches to, work related to AI
    - Is it possible to build an artificial general intelligence (AGI) and/or transformative AI (TAI) system? Is humanity likely to do so?
    - What form(s) is TAI likely to take? What are the implications of that? (E.g., AGI agents vs comprehensive AI services)
    - What will the timeline of AI developments be?
        - How “hard” are various AI developments?
        - How much “effort” will go into various AI developments?
        - How discontinuous will AI development be?
            - Will development *to* human-level AI be discontinuous? How much so?
            - Will development *from* human-level AI be discontinuous? How much so?
            - Will there be a hardware overhang? How much would that change things?
        - How important are individual insights and “lumpy” developments?
        - Will we know when TAI is coming soon? How far in advance? How confidently?
        - What are the relevant past trends? To what extent should we expect them to continue?
    - How much should longtermists’ prioritise AI?
        - How high is existential risk from AI?
        - How “hard” is AI safety?
            - How “hard” are non-impossible technical problems in general?
            - To what extent can we infer that the problem is hard from failure or challenges thus far?
        - Should we expect people to handle AI safety and governance issues adequately without longtermist intervention?
            - To what extent will “safety” problems be solved simply in order to increase “capability” or “economic usefulness”?

                - Would there be clearer evidence of AI risk in future, if it's indeed quite risky? Will that lead to better behaviours regarding AI safety and governance?
            - Could AI pose suffering risks? Is it the most likely source of such risks?
            - How likely are positive or negative "non-existential trajectory changes" as a result of AI-related events? To what extent does that mean longtermists should prioritise AI?
        - What forms might an AI catastrophe take? How likely is each?
        - What are the best approaches to reducing AI risk or increasing AI benefits?
            - From a longtermist perspective, how valuable are approaches focused on relatively "near-term" or "less extreme" issues?
            - What downside risks might (various forms of) work to reduce AI risk have? How big are those downside risks?
                - How likely is it that (various forms of) work to reduce AI risk would accelerate the development of AI? Would that increase overall existential risk?
            - How important is AI governance/strategy/policy work? Which types are most important, and why?
- Value of, and best approaches to, work related to biorisk[10] and biotechnology
    - What will the timeline of biotech developments be?
        - How "hard" are various biotech developments?
        - How much "effort" will go into various biotech developments?
    - How much should longtermists' prioritise biorisk and biotech?
        - How high is existential risk from pandemics involving synthetic biology?
            - Should we be more concerned about accidental or deliberate creation of dangerous pathogens? Should we be more concerned about accidental or deliberate release? What kinds of actors should we be most concerned about?
        - How high is existential risk from naturally arising pandemics?
            - To what extent does the usual "natural risks must be low" argument apply to natural pandemics?
        - What can we (currently) learn from previous pandemics, near misses, etc.?
        - How high is the risk from antimicrobial resistance?
    - How much overlap is there between approaches focused on natural vs. anthropogenic pandemics, "regular" vs. "extreme" risks, etc.?
    - What are the best approaches to reducing biorisk?

[10] Biorisk includes both natural pandemics and pandemics involving synthetic biology. Thus, this risk does not completely belong in the section on "emerging technologies". We include it here anyway because anthropogenic biorisk will be our main focus in this section, given that it's the main focus of the longtermist community and that there are strong arguments that it poses far greater existential risk than natural pandemics do (see e.g. *The Precipice*).

            - What downside risks might (various forms of) work to reduce biorisk have? How big are those downside risks?
- Value of, and best approaches to, work related to nanotechnology
    - What will the timeline of nanotech developments be?
        - How “hard” are various nanotech developments?
        - How much “effort” will go into various nanotech developments?
    - How high is the existential risk from nanotech?
    - What are the best approaches to reducing risks from nanotechnology?
        - What downside risks might (various forms of) work to reduce risks from nanotech have? How big are those downside risks?
- Value of, and best approaches to, work related to interactions and convergences between different emerging technologies

Questions about specific existential risks (which weren’t covered above)

See here for notes and links related to these topics.

- Value of, and best approaches to, work related to nuclear weapons
    - How high is the existential risk from nuclear weapons?
        - How likely are various types of nuclear war?
            - What countries would most likely be involved in a nuclear war?
            - How many weapons would likely be used in a nuclear war?
            - How likely is counterforce vs. countervalue targeting?
            - How likely are accidental launches?
            - How likely is escalation from accidental launch to nuclear war?
        - How likely are various severities of nuclear winter (given a certain type and severity of nuclear war)?
        - What would be the impacts of various severities of nuclear winter?
- Value of, and best approaches to, work related to climate change
    - How high is the existential risk from climate change itself (not from geoengineering)?
        - How much climate change is likely to occur?
        - What would be the impacts of various levels of climate change?
        - How likely are various mechanisms for runaway/extreme climate change?
    - How tractable and risky are various forms of geoengineering?
        - How likely is it that risky geoengineering could be unilaterally implemented?
    - How much does climate change increase other existential risks?
- Value of, and best approaches to, work related to totalitarianism and dystopias
    - How high is the existential risk from totalitarianism and dystopias?
        - How likely is the rise of a global totalitarian or dystopian regime?

        - How likely is it that a global totalitarian or dystopian regime that arose would last long enough to represent or cause an existential catastrophe?
    - Which political changes could increase or decrease existential risks from totalitarianism and dystopia? By how much? What other effects would those political changes have on the long-term future?
        - Would various shifts towards world government or global political cohesion increase risks from totalitarianism and dystopia? By how much? Would those shifts reduce other risks?
        - Would enhanced or centralised state power increase risks from totalitarianism and dystopia? By how much? Would it reduce other risks?
    - Which technological changes could increase or decrease existential risks from totalitarianism and dystopia? By how much? What other effects would those political changes have on the long-term future?
        - Would further development or deployment of surveillance technology increase risks from totalitarianism and dystopia? By how much? Would it reduce other risks?
        - Would further development or deployment of AI for police or military purposes increase risks from totalitarianism and dystopia? By how much? Would it reduce other risks?
        - Would further development or deployment of genetic engineering increase risks from totalitarianism and dystopia? By how much? Would it reduce other risks?
        - Would further development or deployment of other technologies for influencing/controlling people's values increase risks from totalitarianism and dystopia? By how much?
        - Would further development or deployment of life extension technologies increase risks from totalitarianism and dystopia? By how much?

Questions about non-specific risks, existential risk factors, or existential security factors

See here for notes and links related to these topics.

- Value of, and best approaches to, work related to global catastrophes and/or civilizational collapse
    - How much should we be concerned by possible concurrence, combinations, or cascades of catastrophes?
    - How much worse in expectation would a global catastrophe make our long-term trajectory?
        - How effectively, if at all, would a global catastrophe serve as a warning shot?

        - What can we (currently) learn from previous global catastrophes (or things that came close to being global catastrophes)?
    - How likely is collapse, given various intensities of catastrophe?
        - How resilient is society?
    - How likely would a collapse make each of the following outcomes: Extinction; permanent stagnation; recurrent collapse; "scarred" recovery; full recovery?
        - What's the minimum viable human population (from the perspective of genetic diversity)?
        - How likely is economic and technological recovery from collapse?
            - What population size is required for economic specialisation, technological development, etc.?
        - Might we have a "scarred" recovery, in which our long-term trajectory remains worse in expectation despite economic and technological recovery? How important is this possibility?
        - What can we (currently) learn from previous collapses of specific societies, or near-collapses?
    - What are the best approaches for improving mitigation of, resilience to, and recovery from global catastrophes and/or collapse (rather than preventing them)? How valuable are these approaches?
        - (How much) Should we worry about "moral hazard"?
        - (How much) Should we worry about "which world gets saved"?
- Value of, and best approaches to, work related to war
    - By how much does the possibility of various types of wars raise total existential risk?
        - How likely are wars of various levels/types of wars?
            - How likely are "great power wars"?
        - By how much do wars of various levels/types increase existential risk?
            - By how much do great power wars increase existential risk?
- Value of, and best approaches to, work related to improving institutions and/or decision-making
- Value of, and best approaches to, work related to existential security and the Long Reflection
    - Can we achieve existential security? How?
    - Are there downsides to pursuing existential security? If so, how large are they?
    - How important is it that we have a Long Reflection process? What should such a process involve? How can we best prepare for and set up such a process?

We have also collected here some questions that seem less important, or where it's not clear that there's really disagreement on them that fuels differences in strategic views and choices among longtermists. These include questions about "natural" risks (other than "natural" pandemics, which some of the above questions already addressed).

Directions for future work

We'll soon publish a post discussing in more depth the topic of optimal timing for work and donations. We'd also be excited to see future work which:

- Provides that sort of more detailed discussion for other topics raised in this post
- Attempts to actually answer some of these questions, or to at least provide relevant arguments, evidence, etc.
- Identifies additional crucial questions
- Highlights additional relevant references
- Further discusses how beliefs about these questions empirically do and/or logically should relate to each other and to strategic views and choices
    - This could potentially be visually "mapped", perhaps with a similar style to that used in this post.
    - This could also include expert elicitation or other systematic collection of data on actual beliefs and decisions. That would also have the separate benefit of providing one "outside view", which could be used as input into what one "should" believe about these questions.
- Attempts to build formal models of what one should believe or do, or how the future is likely to go, based on various beliefs about these questions
    - Ideally, it would be possible for readers to provide their own inputs and see what the results "should" be

Such work could be done as standalone outputs, or simply by making commenting on this post or the linked Google docs. Please also feel free to get in touch with us if you are looking to do any of the types of work listed above.

This post and the associated documents were based in part on ideas and earlier writings by Justin Shovelain and David Kristoffersson, and with input from them. We also received useful comments from Arden Koehler, Denis Drescher, and Gavin Taylor. Finally, we're grateful to Jesse Liptrap for work on an earlier draft, and to Siebe Rozendal for comments on another earlier draft. This does not imply these people's endorsement of all aspects of this post.