Politics, Policy, and Security from a Broad Longtermist Perspective

A Preliminary Research Agenda

Michael Aird, Staff Researcher at Rethink Priorities, michael@rethinkpriorities.org

Document last updated: 3 October 2021

This document is an early draft. It should not be taken as necessarily representing the views of anyone except me (Michael Aird). *I'm probably happy for you to share this document, but please check with me first.*

Feel free to make comments and suggestions! *Any type of comment/suggestion can be useful, but it'd perhaps be most useful if you say what you feel is "missing" and/or what aspects of my thinking/priorities seem most importantly mistaken. You could also complete this survey or reach out to me for a call.*

Summary

This post outlines a **preliminary** research agenda that Rethink Priorities' (RP's) longtermist team **might** pursue in the future and that we'd be excited to see other people work on as well. This document was written by Michael Aird, who would likely lead RP's work on this agenda.

The topics that might be explored as part of this agenda include:

- Armed conflict and military technology
- Global cooperation and international relations
- Creating and improving institutions and policies
- Safeguarding, strengthening, and/or spreading democracy[1]
- Authoritarianism[2] and/or dystopias[3]
- Risks and opportunities related to China[4]

The primary reasons for grouping these topics together are practical ones:

- **There's substantial overlap in the knowledge, skills, and connections** that would help a person research - and influence important decisions related to - each topic in this group
- Things we learn and conclude while exploring each topic might **inform our research or recommendations on other topics in this group**

Meanwhile, the primary reasons for RP to potentially work on these topics are that:

- These topics contain many questions that seem **neglected and important**

---

[1] It's worth noting that I'm not assuming that democracy - or actions aimed at safeguarding, strengthening, or spreading it - will always or entirely be good things. In fact, one of the questions that might be tackled as part of this agenda is about potential downsides of those things. For example, increasing the public's influence on policy-making may sometimes lead to more short-termist, ill-informed, or incoherent policies. For another example, efforts by some nations to impose regime change on others - or mere rhetoric or fears about such regime change efforts - could cause armed conflict, cause instability, or incentivise the development of WMDs.

[2] I (Michael) have the impression that longtermists have often focused more on totalitarianism than authoritarianism, or have used the terms as if they were somewhat interchangeable. But my understanding is that political scientists typically consider totalitarianism to be a relatively extreme subtype of authoritarianism (see, e.g., Wikipedia). And it's not obvious to me that, from a longtermist perspective, totalitarianism is a bigger issue than other types of authoritarian regime. (Essentially, I'd guess that stable totalitarianism would have worse effects than other types of stable authoritarianism, but that it's less likely to arise in the first place.) As such, this document uses the inclusive term "authoritarianism".

[3] By "dystopia", I essentially have in mind what Ord (2020) refers to as an "unrecoverable dystopia": a type of existential catastrophe in which "civilization [is] intact, but locked into a terrible form, with little or no value". Ord lists three subtypes:

- Undesired dystopia (e.g., a world in which relentless market or genetic competition drives us towards maximum efficiency or fitness, at the cost of things we value)
- Enforced dystopia (e.g., a global, stable totalitarian regime)
- Desired dystopia (e.g., "worlds that forever fail to recognise some key form of injustice [and] thus perpetuate it blindly")

[4] Many of the subtopics and questions in this section of the agenda could also productively be asked regarding other countries or regions, perhaps most notably Russia.

- There seems to be **no or very few organisations** with a substantial focus on researching these topics from a longtermist perspective in a relatively broad way (e.g., without primarily focusing on the intersection of these topics with AI)
- RP may have a **comparative advantage** for doing that sort of research

That said, grouping these topics together and prioritising them could also be motivated by adopting a **broad longtermist** approach, a focus on **existential risk and security factors**, and/or a focus on **non-extinction existential risks** (including some suffering risks).

There would be different theories of change for different parts of this agenda, as well as arguments against pursuing parts of this agenda; these points are discussed below.

Why you might want to read this agenda

See also Potential benefits & downsides of making and/or sharing a research agenda.

- **Researchers or potential researchers** might want to read this agenda to:
    - Help them decide whether to express interest in working on this agenda with RP as a volunteer, collaborator, intern, or permanent hire.
    - Find questions to pursue independently of RP.
        - This agenda contains far too many questions for RP to tackle them all, and many other people and organisations will have a comparative advantage for specific questions.
- **Funders (including small donors)** might want to read this agenda to:
    - Help them decide whether to provide RP with funding to start on this research agenda earlier and make faster progress on it (by hiring additional interns or permanent hires with relevant skills and interests).
- **A wide range of people** might want to read this agenda to see a structured collection of many potentially important and under-discussed uncertainties, problems, interventions, and considerations, along with a rationale for some people to focus on these things.
    - This might be helpful in a range of ways. For example:
        - This might help researchers generate, evaluate, and pursue research questions beyond those listed in this agenda.
        - This might help funders seek out and evaluate new funding opportunities.
        - This might help policy makers, advisors, and advocates generate, evaluate, optimise, and argue for policy ideas.
        - This might help people generate and evaluate options for their own careers.
- Finally, **a wide range of people** might also want to read this agenda so they can provide feedback on it (e.g., using this survey) and thereby help us have a positive impact.

Each section of the agenda can be read independently, and some people may wish to read only certain sections rather than the whole thing.

Table of contents

Background and motivation

How does this research agenda fit into RP's broader plans?

This research agenda is intended to be one part of RP's longtermist work. By default, **we would likely commence work on this agenda in 2022**, after we finish our project on nuclear risk. However, given sufficient funding and interested applicants, it's possible an intern or a new hire would start on this agenda before then.

Work on this agenda would both complement and be complemented by other ongoing or planned work at RP, such that it would be beneficial to house these different projects in the same organisation. This includes some of RP's work in cause areas other than longtermism, such as our cross-cause or animal-focused work on polling, ballot initiatives, and politics and policymaking in the US, the EU, and China (see our list of publications). These points are discussed further in a later section.

For an overview of RP's plans for 2021 and general theory of change, see Davis (2020).

Why should *someone* do research related to broad longtermism and existential risk & security factors?

Todd (2020) highlights four main "varieties of longtermism":

- **Patient longtermism:** The view that one should focus on making it more likely that good decisions are made more than a few decades from now
- **Broad urgent longtermism:** The view that one should focus on making it more likely that good decisions are made within the coming decades, but with it being unclear which sorts of decisions (e.g., decisions on AI vs biorisk vs electoral systems vs space governance) are particularly important (see also "Broad vs. narrow interventions")
- **Targeted urgent longtermism focused on existential risks:** The view that one should focus on making it more likely, in the coming decades, that good decisions are made in relation to one or more of a small set of existential risks
- **Targeted urgent longtermism focused on other trajectory changes:** The view that one should focus on making it more likely that, in the coming decades, good decisions are made in relation to one or more of a small set of things that could affect the long-term future in ways *other than* by causing or preventing existential risks (e.g., ensuring we capture 80% rather than 79% of the possible value of the future)

For the purposes of this document, the most important distinction is the distinction between broad urgent longtermism and targeted urgent longtermism focused on existential risks. Broad urgent longtermism can push in favour of a focus on existential risk factors and existential security factors: things that make one or more existential catastrophes more or less likely, without themselves directly causing or preventing existential catastrophes. Examples could include economic stagnation, war, and a spreading or strengthening of democracy (see Ord,

2020).[5] In contrast, targeted urgent longtermism focused on existential risks would focus on a specific set of things that could directly cause or prevent existential catastrophe, such as (perhaps) artificial intelligence or nuclear war.

In my (Michael's) view, it's currently very hard to say how much weight longtermists as a whole should give to each of the four above-mentioned perspectives, such that:

1. All should get substantial weight
2. Decisions to specialise for work on one perspective or another should probably focus more on how neglected each perspective is and what a person's comparative advantage is, rather than how much weight longtermists as a whole should give each perspective[6]

Currently, I have the impression that longtermists as a whole are effectively giving substantially more weight to targeted urgent longtermism focused on existential risks than to the other perspectives. Relatedly, many questions that are high priority from the other perspectives remain entirely ignored or only very superficially explored. Furthermore, there currently seems to be no organisation that has more than a couple researchers who (a) are motivated by longtermism and (b) collectively focus on a range of topics that are high priority from a broad longtermist perspective (without restricting their scope to just one particular field or topic).[7]

---

[5] By the same token, broad urgent longtermism can also push in favour of a focus on risk and security factors for s-risks. And in any case, many factors may be risk or security factors for *both* s-risks and other existential risks.

[6] This is related to the idea of taking "the portfolio approach", 80,000 Hours' discussion of comparative advantage, and Tomasik's discussion of Why Charities Usually Don't Differ Astronomically in Expected Cost-Effectiveness. See also Baumann (2017).

All this being said, I still do think that one useful question to ask when setting one's individual priorities is "What is my current best guess as to how much weight longtermists as a whole should give to each perspective?" (This could be asked alongside questions about neglectedness, comparative advantage, which actions provide the most information about one's comparative advantage, etc.) Additionally, I think that further research that reduces our uncertainty about that could be highly valuable.

[7] This isn't intended as a criticism of existing organisations; I don't think that all organisations should fit that description, but rather that there should be at least one or a few organisations that do fit that description.

Organisations that come relatively close to fitting these description include:

- Founders Pledge (but they only have a couple researchers focused on broad longtermist topics, those researchers may only conduct relatively brief investigations and then move on to other topics, and their research focuses on identifying funding opportunities rather than also on other paths to impact)
- The Improving Institutional Decision-Making working group (but that's a volunteer collaboration rather than an organisation, they aren't necessarily primarily motivated by longtermism, they have a substantial emphasis on non-research activities, and they're focused only on improving institutional decision-making)
- The Centre for the Governance of AI (but that's quite focused on AI)
- The Center for Security and Emerging Technology (but they aren't necessarily primarily motivated by longtermism, and they have a substantial emphasis on non-research activities and on the topic of AI in particular)
- The Simon Institute for Longterm Governance (but that currently has only two staff members, they have a substantial emphasis on non-research activities, and they're mostly or entirely focused just on improving institutions and policies)

Similar points also seem true in relation to existential risk and security factors.

(All that said, similar points might also be true in relation to patient longtermism and/or targeted longtermism focused on other trajectory changes. Therefore, as alluded to above, I expect there should also be some people specialising for work on one of those perspectives. But as discussed in a later section, RP seems to have a comparative advantage for research that's particularly relevant to a broad longtermist perspective.)

Why should *someone* do research related to non-extinction existential risks?

My answer to this question mirrors the answer I gave above:

- In my view, we're currently very uncertain about how the likelihood and tractability of extinction and non-extinction existential risks compare, such that:
    - Both categories of risks should get substantial attention
    - Decisions to specialise for work on one category of risks or the other should probably focus more on how neglected each category is and what one's comparative advantage is, rather than how likely and tractable each category of risks is
- There appears to be a substantially larger amount of rigorous work done and planned on extinction risk than on non-extinction existential risk
    - Perhaps especially when it comes to the risk of an unrecoverable dystopia, rather than the risk of an unrecoverable collapse

Why should *Rethink Priorities* do research related to broad longtermism, existential risk & security factors, and/or non-extinction existential risks?

There are five main reasons why RP in particular is suited to doing this research.

First, RP's longtermism-focused researchers are relatively early in their research careers and in their time with RP. As such, neither these researchers as individuals nor RP's longtermism team as an entity have specialised to a particularly strong extent yet. Thus, RP may currently be **well-positioned to specialise to fill any given neglected niche**. And as argued above, research related to broad longtermism, existential risk and security factors, and/or non-extinction existential risks may be one such niche.

There are also many non-longtermist organisations and individuals working on the same sorts of topics as those which seem high priority from a broad urgent longtermist perspective. For example, many non-longtermists work on safeguarding, strengthening, or spreading democracy; topics related to politics and technology in China; and global cooperation and peacebuilding. But these people tend to overlook many specific questions that longtermists would consider particularly important.

Second, relative to other effective-altruism- or longtermism-aligned research organisations, RP has tended to place somewhat more emphasis on **empirical work, quantitative work, and syntheses of existing work that was conducted by people focused on somewhat different questions**, relative to theoretical, abstract, or entirely novel work. For example, RP's previous work on nuclear weapons involved relatively little development of theories or concepts, and relatively more synthesis of a wide array of empirical data and construction of quantitative models. Arguably, RP has built up a degree of comparative advantage for the use of these approaches. And arguably these approaches are particularly useful for many areas that are of more interest to broad rather than targeted longtermism, due to those areas having a greater availability of empirical data and existing non-longtermist work.

Third, several RP staff members, and RP as an organisation, have already built up some degree of **knowledge or skills relevant to the specific topics included in this research agenda**. For example:

- RP's work on nuclear war has included work related to armed conflict and military technology; global cooperation and international relations; creating and improving institutions and policies; and authoritarian regimes such as Russia, North Korea, and China
- RP's animal welfare work has included work on policymaking, the EU, and China
- RP has done politics- and policy-related polling and message testing
- RP has done work related to deliberative democracy, ballot initiatives, forecasting (which could be used for improving institutional decision-making), and electoral reform
- Two RP staff members (Neil Dullaghan and Dominika Krupocin) have PhDs in political science and security studies, respectively
- One RP staff member (Linch Zhang) is proficient in Mandarin
- Linch Zhang and I have each previously done (non-public) work related to democracy, authoritarianism, and/or dystopias

Fourth, there would be **complementarities between the work outlined in this research agenda and other ongoing or planned work at RP**, as described earlier.

Fifth, some stakeholders and advisors have indicated in conversation that **they would be excited to see more longtermist work on some of these topics, including from RP in particular**.

Why group these topics together?

There are three main reasons to group these topics together.

First, there's **substantial overlap in the knowledge, skills, and connections** that would help a person research - and influence important decisions related to - each topic in this group. For example:

- Knowledge from the fields of international relations, political science, and history - and knowledge of the methodologies used in those fields - would be relevant to all of these topics
- Knowledge about authoritarianism in general is helpful in understanding risks and opportunities related to China, and vice versa

Second, things we learn and conclude while exploring each topic might **inform our research or recommendations on other topics in this group**. For example:

- Research on armed conflict and military technology should give indications about (a) how valuable various international relations efforts, institutions, and policies would be and (b) what key factors one should consider when doing research to prioritise or inform such efforts
- Research on the benefits, harms, and durability of democracy and authoritarianism should yield insights about what the costs and benefits of various types of armed conflict, military technology, and international relations efforts would be[8]

Third, as noted earlier, work on each of these topics can be **motivated by similar rationales**, such as:

- Adopting a broad longtermist approach
- Adopting a focus on existential risk and security factors
- Adopting a focus on non-extinction existential risks

What would be the theories of change for this work?

This preliminary agenda covers a vast array of topics and methodologies, and we haven't yet spent much time operationalising the questions or working out what to (de)prioritise. Thus, different parts of the agenda would have quite different theories of change, and they would be fleshed out further as we begin exploring or tackling specific parts of this agenda. But here are a few general points that can be made already. (Elaboration and other relevant points can be found in Why EAs researching mainstream topics can be useful.)

- Work on this agenda would be aimed at gaining clarity on either how much a particular topic matters for the long-term future, the precise pathways and mechanisms by which it matters for the long-term future, or what to do about it
    - For example, gaining more clarity on how important the issue of authoritarianism in general is, how bad various types of authoritarianism would be and precisely why, and what the most cost-effective interventions for dealing with these issues are
- Some work on this agenda would focus simply on **bringing existing knowledge, theories, etc. into the EA or longtermist communities**, and helping decision-makers in those communities make decisions in light of those things

[8] For example, the more we should worry about stable, global authoritarianism relative to other existential risks, the less obvious it is that longtermists should support more intensive forms of global cooperation or reductions in liberal democracies' ability and willingness to engage in armed conflict.

- Other work on this agenda would tackle **questions that are important for the long-term future and that differ from those tackled in existing work on a topic** - particularly questions focused on implications for the long-term future, but also questions focused on cost-effectiveness and probabilistic forecasts
- The types of decisions we would focus on influencing include decisions about **funding, policy, other research, and careers**, with funding as perhaps the primary focus
- We expect to mostly (a) influence the decisions of actors with some connection to the longtermist community, or (b) *via* such actors, indirectly influence actors with no connection to the longtermist community
    - But we would also aim to have *some* direct influence on actors with no connection to the longtermist community

Arguments against pursuing this agenda

This section just briefly highlights some potential arguments *against* pursuing this agenda. I would be happy to, in the comments section, elaborate on these arguments, counterpoints to them, and what implications I think they have for work on this agenda.

1. **Importance**: Many of these questions may have little relevance to decisions that are key for the long-term future.
2. **Tractability**: Many of these questions may be hard to make any progress on. It may also be hard to influence many of the decision-makers who are best-positioned to act on the information this work could provide (e.g., the US government, the UN, elites in China)
3. **Neglectedness**: Many of these questions, or similar questions, have already been researched to some extent by other longtermists or non-longtermists, or can be expected to be in future.
4. **Comparative advantage**: One could argue that the knowledge, skills, credentials, and connections of I or others at RP are poorly suited to work on this agenda or better suited to other high-priority work.
5. **Opportunity cost**: Even if work on this agenda is important, tractable, neglected, and RP's comparative advantage, RP might have even more impact by spending the time that would've been spent on this agenda on other topics instead.
6. **Downside risks**: Some work on this agenda could pose information hazards or other downside risks, at least if it's framed and shared in certain ways.

Overall, I think RP should devote some staff time to this agenda despite these arguments, partly because we could plan and conduct our research with these arguments in mind. (See also Why EAs researching mainstream topics can be useful.)

Some potential topics, subtopics, and questions

What follows is a preliminary list of topics, subtopics, and specific questions that might be covered as part of this research agenda.

This list should be taken as **illustrative rather than definitive**. As noted earlier, barring possible work by interns, we're unlikely to work on this agenda until we wrap up our work on nuclear risk or receive sufficient funding to hire an additional researcher with relevant skills and interests. Thus, we haven't yet spent a lot of time deciding precisely what topics and questions this agenda should and shouldn't cover or how best to categorise them, or precisely operationalising these questions, determining theories of change for them, and working out what to (de)prioritise. For example, in practice, we might often investigate much narrower versions fo the questions shown below, such as versions that focus on just a handful of specific technologies, countries, or historical case studies.

Additionally, we have some further thoughts on these research ideas - and some additional research ideas - that are not noted here. We also haven't yet explicitly indicated which academic fields are relevant to which questions, but we may do so in future.

Armed conflict and military technology

- How likely are international tensions, armed conflicts of various levels/types, and great power war specifically at various future times? What are the causes of these things?
    - How often do international tensions escalate into armed conflicts? How often do armed conflicts of various types result in (substantial) vertical, horizontal, and/or political escalation? What causes or prevents such escalation?
    - What have the historical trends in these likelihoods been?
    - How are these likelihoods expected to change in future?
    - How have the causes changed over time?
    - What will the causes be in future?
    - How might plausible changes in variables such as technology, climate, power, resource scarcity, migration, urbanisation, population size, and economic growth affect answers to the above questions?
        - To what extent does this push in favour of or against work to affect those variables (e.g., climate change mitigation, open borders advocacy, improving macroeconomic policy)?
    - Are Pinker's claims in *The Better Angels of Our Nature* essentially correct?
    - Are the current trends likely to hold in future? What might affect them?
- How much, and in which direction, do international tensions, strategic competition, and risks of armed conflict affect the expected value of the long-term future? By what pathways?[9]
    - (Obviously the above questions about likelihoods and causes are relevant here as well)
    - What are the plausible ways a great power war could play out?

[9] This subtopic can be seen as elaborating on some questions in the post Crucial questions for longtermists.

        - E.g., what countries would become involved? How much would it escalate? How long would it last? What types of technologies might be developed and/or used during it?
    - What are the main pathways by which international tensions, armed conflicts of various levels/types, or great power war specifically could increase (or decrease) existential risks? Possible examples include:
        - Spurring dangerous development and/or deployment of new technologies
        - Spurring dangerous deployment of existing technologies
        - Impeding existential risk reduction efforts (since those often require coordination and are global public goods)
        - Sweeping aside or ushering in global governance arrangements
        - Weakening (or strengthening) democracies
        - Worsening (or improving) the values of various actors (e.g., reducing or increasing impartiality or inclinations towards multilateralism among the public or among political leaders)
        - Changing the international system's global governance arrangements and/or polarity (which could then make coordination easier or harder, make stable authoritarianism more or less likely, etc.)
        - Serving as a "warning shot" that improves values, facilitates coordination, motivates risk reduction efforts, etc.
    - All things considered, by how much do international tensions, strategic competition, armed conflicts of various levels/types, and great power war specifically increase (or decrease) existential risk?
    - To what extent, and by what pathways, would international tensions, strategic competition, armed conflicts of various levels/types, and great power war specifically reduce the expected value of the long-term future in ways unrelated to existential risk?
        - E.g., slowing down progress, increasing the chance of especially terrible rather than "merely bad" futures, or reducing the chance of especially excellent rather than "merely good" futures.
    - How will new technologies affect answers to the above questions?
        - Technologies worth considering include advanced AI, advanced biotechnology, and autonomous weapon systems.
        - Relevant implications of new technologies could include increasing or decreasing strategic stability, increasing or decreasing polarity, and increasing or decreasing the chance that armed conflict would lead to existential catastrophe.
    - How might plausible changes in variables such as technology, climate, power, resource scarcity, migration, urbanisation, population size, and economic growth affect answers to the above questions?
- What are the best actions for intervening on international tensions, strategic competition, risks of armed conflict, or specifically the ways that these things might harm the long-term future?

- Here and for all other questions in this research agenda about “best actions”, we can ask the following subquestions:
  - What organisations are working on these issues?
    - What can we learn from them?
    - How effective do they seem to be?
    - How much good could they do with additional funding, talent, or advice?
    - In what situations would providing such funding, talent, or advice actually be bad for the world?
  - What actions have most successfully achieved those goals in the past? What made those actions so successful? Would the same or similar actions still be effective in present and future contexts?
  - What are the most cost-effective actions for achieving these goals?
  - What are the actions that achieve the greatest effect per talented person focused on them?
  - What are the actions that longtermists have a comparative advantage for?
  - What commonly proposed or seemingly good actions are counterproductive or inefficient?
- In relation to international tensions, strategic competition, and risks of armed conflict in particular, we can also ask the following specific sub-questions:
  - How useful are things like diplomacy, treaties, arms control agreements, international organisations, and international norms? What actions are best in relation to those things?
    - What are the best levers for influencing existing or future arms control regimes? E.g., public advocacy vs targeted advocacy vs technical support; new treaties vs amendments to or expansion of existing treaties vs better implementation of existing treaties vs export control regimes; influencing great powers vs middle or smaller powers vs international organisations vs non-state actors?
    - What was the relative importance of academic research, non-profit advocacy, and diplomatic work for achieving past arms control agreements?
    - What should longtermists do in light of that (e.g., should more move into careers in those areas)?
    - How much influence can treaties have in cases where compliance is hard/impossible to verify or enforce, and/or where relevant actors don’t become parties to the treaty? What factors affect that? What are the relevant mechanisms for influence?
  - Should our actions in this area focus on international tensions, armed conflict in general, or specific types of armed conflict (e.g., great power war)?
  - Should our actions in this area focus on preventing these things from occurring at all, preventing them from escalating, or countering specific ways in which they could harm the future (e.g., reducing the chance that,

conditional on a great power war occuring, dangerous technologies are developed and deployed)?
        - Are there ways of carving up the space of possible focuses that are more useful than the ways used in the above two bullet points?
- To what extent, and in what ways, is research, development, production, and proliferation of militarily relevant technologies affected by self-interested actions by actors such as corporations and legislators?
    - This topic is intended to cover:
        - Actions related to the influence of the defence industry, pork-barrel politics, and regulatory capture. (See, for example, this discussion of how such things may influence the US's stance on its ICBMs.)
        - Effects such as arms races, development of technologies that undermine deterrence, development of technologies that could pose existential risk or are prerequisites to technologies that could, or proliferation of such technologies.
    - To what extent, and in what ways, have such actions caused such effects in the past?
    - To what extent, and in what ways, will such actions cause such effects in the future?
    - To what extent do such actions thereby affect existential risks? By what pathways?
    - What are the best actions for intervening on these issues?
- Governance to mitigate risks from research, development, production, and (mis)use of militarily relevant technologies[10]
    - What relevant governance efforts have been successful or unsuccessful in the past? What can we learn from those examples?
        - Relevant governance efforts could be national or international, and include include technological regulations, policies regarding dual-use research of concern, arms control agreements, export control regimes
    - What relevant governance efforts are currently being used or advocated?
    - What relevant governance efforts might be useful in the present or the future? What relevant efforts might *not* be useful?
        - For example, is the idea of "AI arms control" useful, and what AI arms control efforts might be worth pursuing?
- Public attitudes, social norms, and taboos related to armed conflict, militarily relevant technology, and specifically WMDs[11]

[10] This subtopic is inspired in part by a question from the post Research questions that could have a big social impact, organised by discipline and some ideas in the post Project Ideas in Biosecurity for EAs.
[11] This subtopic is inspired in part by an idea in the post Project Ideas in Biosecurity for EAs.

Relevant public attitudes, social norms, and taboos could include those related to:
- engaging in armed conflict in general
- preparing for armed conflict in general
- engaging in or preparing for specific types or methods of armed conflict
- engaging in research, development, or production of specific militarily relevant technologies

    - How much and in what ways have public attitudes, social norms, and taboos each affected armed conflict in the past? How much and in what ways are they likely to affect armed conflict in the future? By what specific mechanisms?[12]
    - How much and in what ways have public attitudes, social norms, and taboos each affected research, development, production, and use of militarily relevant technology (and especially WMDs) in the past? How are they likely to affect these things in the future? By what specific mechanisms?
        - E.g., how much have norms and taboos protected against especially risky development or use of biotechnology?
        - E.g., is it plausible that a taboo against large-scale development or use of autonomous weapons systems could emerge?
    - What are the drivers of relevant public attitudes, social norms, and taboos? How durable or fragile are they? To what extent do they tend to adapt well to changing circumstances, remain overly fixed, or drift towards counterproductive or overly simplistic forms[13]?
    - How well can thoughtful actors influence the emergence or direction of relevant public attitudes, social norms, and taboos?
    - How often (if ever) are relevant public attitudes, social norms, and taboos counterproductive by creating divides between groups (e.g., civil society and the military) or causing some groups to see others as alarmist and naive?
    - What are the best actions for influencing relevant public attitudes, norms, and taboos?
    - How would the answers to all of the above questions differ across different countries, cultures, types of armed conflict, types of militarily relevant technology, etc.?
- Non-state actors and WMDs[14]
    - How many non-state actors have there been whose motivations and capabilities made it plausible that they would develop and/or use WMDs?
    - What has prevented these non-state actors from actually developing and/or using WMDs?
    - What has prevented there from being more non-state actors with particularly worrying motivations and capabilities?
    - What have been the more specific motivations of relevant non-state actors?
        - E.g., what specific types of WMDs did they want to use, in what ways, and for what purposes?

- engaging in use - or *specific types of* use - of specific militarily relevant technologies

A prominent example of such a taboo is the purported nuclear taboo.

[12] Relevant mechanisms could include internalisation by political leaders, internalisation by military officials, internalisation by civil society actors who then pressure leaders in targeted ways, and internalisation by segments of the public who then pressure leaders via things like marches and elections.

[13] For some example prior discussion of similar topics, see Gentzel (2017a, 2017b, 2021).

[14] This subtopic is inspired in part by an idea in the post Project Ideas in Biosecurity for EAs. Note that we could also ask all the same questions with respect to state actors, and especially small or developing state actors, for example North Korea.

- Autonomous weapons systems (AWSs)
    - Do AWSs increase existential risk? If so, by how much, and by what mechanisms?
    - What are the best actions for reducing risks from AWSs?
    - Are claims from the Future of Life Institute and Anthony Aguirre on the above questions essentially correct?
    - What are the benefits, risks, and costs of the Campaign to Stop Killer Robots? Has it been a good use of resources? Should it absorb more resources?
    - Would a treaty restricting the development or use of lethal autonomous weapons be beneficial? Could compliance be reliably verified? If so, how? If not, how influential would such a treaty be?[15]

Global cooperation and international relations

- World government
    - Which forms of world government, or of movements towards world government, are most likely?
    - What are the longtermism-relevant benefits and harms of world government in general, various particular forms of it, or various particular movements towards it? How large are these benefits and harms?
        - E.g., to what extent and in what ways might particular forms of world government decrease armed conflict, decrease non-state use of WMDs, or increase the risk of stable, global authoritarianism?
        - E.g., how persistent would a world government be? How much would its initial form influence its later form?
    - What are the best actions for affecting which forms of (movements towards) world government occur?
        - Here we can ask the same sub-questions about "best actions" that were listed earlier
    - Might there be particular future periods when movements towards world government would be particularly likely? What are the implications of this?
        - Would such changes be particularly likely following a great power war or world war?
            - What can we learn about this question from the events that followed previous wars, such as the formation of the League of Nations and the United Nations following World Wars I and II, respectively?
        - Does this suggest we shouldn't focus on this issue unless and until such periods arise or start to seem likelier and closer?
        - Does this mean we should position ourselves to have as much influence as possible if and when such periods arise?
            - What would be the best actions to position ourselves for this?

[15] This question is adapted from a question from the post Research questions that could have a big social impact, organised by discipline.

- Global governance, international institutions, and international cooperation more broadly
    - Equivalent questions to the above questions on world government would also be relevant here

Creating and improving institutions and policies

- How well, and by what processes, do institutions tend to recognise what issues are worth making explicit decisions about, make important decisions, set important policies, and prepare for or respond to catastrophes?[16]
    - What are the drivers of these tendencies?
    - How might this change in future (e.g., given changes in technology, geopolitics, political polarisation, and experiences with past catastrophes such as COVID-19)?
    - What would be the best actions for improving these tendencies?
        - Here we can ask the same sub-questions about "best actions" that were listed earlier
- Longtermism-relevant indices and proxies
    - From a longtermist perspective, which are the most useful indices/proxies for institutions and policies to target or be evaluated against?
    - What would be the benefits and harms of institutions and policies targeting or being evaluated against these indices/proxies, compared to the indices/proxies typically used today (e.g., GDP per capita)?
        - What unintended consequences might occur?
    - What would be the best actions for changing which indices/proxies various institutions and policies target or are evaluated against?
        - What can be learned from previous or ongoing similar efforts, whether motivated by longtermism or not? (For example, efforts to move towards a focus on indicators of wellbeing.)
- Aligning institutions and policies with the interests of future moral patients (potentially including some non-human beings) (see also institutions for future generations)[17]
    - What new institutions would be best for accurately and effectively representing the interests of future moral patients?
    - What new policies or adjustments to existing institutions would be best for accurately and effectively representing the interests of future moral patients?
    - What mechanisms *other than* "representing future moral patients" would be best for aligning institutions and policies with those individuals' interests?
        - E.g., mechanisms centred on better analysis of and preparation for risks (without necessarily explicitly emphasising future moral patients), or on

[16] This subtopic and the next one are adapted from a question from the post Research questions that could have a big social impact, organised by discipline.
[17] This subtopic could relate to subnational, national, or international political institutions and policies, as well as to private sector or civil society institutions.

accounting for externalities affecting *present or near-future* humans and/or non-humans[18]

- What are the best actions for getting good ideas actually turned into good policies and ensuring the policies are implemented properly?
- What are the best actions for increasing (decreasing) the rate at which politicians who'd be relatively good (bad) for the long-term future run for office and are elected? Is it best to focus on (a) increasing the rates for politicians who'd be relatively good or (b) decreasing the rates for politicians who'd be relatively bad? Is it best to focus on the rate for politicians who'd be *especially* good or bad?[19]
- Which policy areas should be the focus for the above questions and efforts?
  - E.g., technology governance, space governance, international cooperation, machine ethics, animal welfare
- Which levels of government should be the focus for the above questions and efforts?
  - Options include supranational (e.g., UN, EU), federal/national, state, and local, but other taxonomies are also possible.

Safeguarding, strengthening, and/or spreading democracy

(Note that, as mentioned in footnote 1, I don't mean to imply that it's definitely, always, or entirely good to safeguard, strengthen, or spread democracy.)

- Ways political systems could be more or less "democratic", and their longtermist implications
  - From a longtermist perspective, what are the potentially significant ways in which a political system could be more or less "democratic", "liberal", etc., and what's a useful taxonomy of these ways?
  - What might the longtermism-relevant implications of a system being more or less democratic in each of those ways be? How good or bad are those implications?
  - Does democratic backsliding have meaningfully similar effects to, or meaningfully increase the likelihood of, authoritarianism or dystopias? Or is it a much less extreme or relevant phenomena?
  - Overall, does a political system being more democratic in each of those ways seem good or bad for the long-term future? To what extent? Under what conditions?
    - E.g., how would the effects of a marginal shift towards "more democracy" differ if it happens in the US vs in India vs in China?

---

[18] Many longtermist priorities relate to transgenerational global public goods (Bostrom, 2013) and/or potential moral catastrophes (William, 2015). These things are currently more neglected than they should be *partly* because many of the beings affected by them would exist in the far future, but also partly because of their global nature (inducing free-rider problems) or because they affect non-humans. It therefore seems the neglectedness of these priorities could be reduced by better accounting for externalities to present or near-future humans (regardless of country) or nonhumans, even without "representing future generations".

[19] See also Reducing long-term risks from malevolent actors.

- Current and expected trends
    - Are current trends towards more or less democracy?
    - What trends should we expect for the future?
    - What's driving these trends? What might affect them in future?
    - How do answers to the above questions differ from place to place (with a particular eye on key places like the US and China)?
    - How do answers to the above questions differ between different dimensions relevant to how "democratic" a place is?
- Interventions
    - Here we can ask the same sub-questions about "best actions" that were listed earlier[20]

Authoritarianism and/or dystopias[21]

- Longtermism-relevant typology and harms of authoritarianism[22]
    - What is the most useful way for longtermists to carve up the space of possible types of authoritarian political systems (or perhaps political systems more broadly, or political systems other than full liberal democracies)? What terms should we be using?
    - Which types of authoritarian political systems should we be most focused on? To what extent should we focus on totalitarian systems?
    - What are the main pathways by which each type of authoritarian political system could reduce (or increase) the expected value of the long-term future?
        - E.g., increasing the rate or severity of armed conflict; reducing the chance that humanity has (something approximating) a successful long reflection; increasing the chances of an unrecoverable dystopia.
    - All things considered, how large does the direct existential risk from global, stable authoritarianism seem to be?
    - All things considered, how large of an existential risk factor does authoritarianism seem to be?
- Risk and security factors for (global, stable) authoritarianism[23]
    - How much would each of the "risk factors for stable totalitarianism" reviewed by Caplan (2008) increase the risk of (global, stable) authoritarianism (if at all)?
    - How likely is the occurrence of each factor?
    - What other risk or security factors should we focus on?
    - What effects would those factors have on important outcomes other than authoritarianism? All things considered, is each factor good or bad for the long-term future?

---

[20] Here the sub-question "What commonly proposed or seemingly good actions are counterproductive or inefficient?" might be especially important.
[21] Research on this topic could also be complemented by the framings and ideas in the post Reducing long-term risks from malevolent actors.
[22] See footnote 2 of this document for some initial thoughts on this.
[23] This subtopic can be seen as elaborating on some questions in the post Crucial questions for longtermists.

        - E.g., mass surveillance, preventive policing, enhanced global governance, and/or world government might be risk factors from the perspective of authoritarianism but security factors from the perspective of extinction or collapse risks (see also Bostrom, 2019).
    - What are the best actions for influencing these factors?
        - Here we can ask the same sub-questions about "best actions" that were listed earlier
    - Can we use changes in these factors as predictors of (stable, global) authoritarianism?
        - Does this suggest we shouldn't focus on this issue unless and until those factors change in concerning ways?
        - Does this mean we should position ourselves to have as much influence as possible if and when those factors start changing in those ways?
            - What would be the best actions to position ourselves for this?
        - Are there other predictors we should also pay attention to?
- How likely is it that relevant kinds of authoritarian regimes will emerge, spread (especially to become global), and/or persist (especially indefinitely)?
    - How politically and technologically feasible would this be?
        - Under what conditions would societies trend towards and/or maintain authoritarianism or a lack thereof?
        - What strategic, military, economic, and political advantages and disadvantages do more authoritarian regimes tend to have? How does this differ based on factors like the nature of the authoritarian regime, the size of the state/polity it governs, and the nature and size of its adversaries?
    - How likely is it that relevant actors will have the right motivations to bring this about?
    - How many current political systems seem to be trending towards authoritarianism?
    - How much (if at all) are existing authoritarian regimes likely to spread? How long are they likely to persist? Why?
    - How likely is it that any existing authoritarian regimes would spread *globally* and/or persist *indefinitely*? Why?
    - How likely was it that any historical authoritarian regimes could've spread much farther and/or lasted much longer than they did? What factors, if any, might've allowed them to spread much farther and/or last much longer than they did? What does this all tell us about future risks from authoritarianism, and how to mitigate such risks?
        - For example, how much would've had to have gone differently for Nazi Germany to win WWII or the Soviet Union to become the dominant world power? And if they had, how long might their dominance have lasted?
- Typology of, likelihoods of, and interventions for dystopias
    - Could an alternative or more detailed typology of dystopias than that presented in Ord (2020) be more useful? What would such a typology look like?

    - How likely is each type of dystopia to arise initially and then to persist indefinitely?
    - How bad would each type of unrecoverable dystopia be, relative to each other, to other existential catastrophes, and to other possible futures?
        - How would the answer change given different plausible moral views, decision theories, or approaches for handling moral and decision-theoretic uncertainty?
    - How much should we worry about *recoverable* or *temporary* equivalents of each type of unrecoverable dystopia?
        - E.g., how much would each increase (or decrease) the risk of later extinction, unrecoverable collapse, or unrecoverable dystopia?
    - What are the main factors affecting the likelihood, severity, and persistence of each type of dystopia?
    - What would be the best actions for reducing the likelihood, severity, or persistence of each type of dystopia?

Risks and opportunities related to China

- Politics and policymaking in China
    - How are policies set in China, especially on topics of relevance to longtermism?
        - How has this changed over time (e.g., before vs after Xi Jinping became the paramount leader)?
        - How might this change in future?
    - How much influence do various actors have on politics and policymaking in China?
        - E.g., officials, elites, the Chinese public, perhaps civil society, perhaps overseas actors
    - How do answers to the above questions vary between different longtermism-relevant policy areas (e.g., AI, biotechnology, pandemic preparedness, surveillance, civil liberties, disaster management)?
- China's attitudes and policies regarding potentially risky technologies[24]
    - What attitudes regarding potentially risky technologies are held by the Chinese government, Chinese elites, the general public in China, etc.?
        - Possible ways to investigate this include looking at policy documents, how various tech companies are regulated, what is included in relevant documentaries shown on state-controlled TV broadcasters, funding levels for various institutions and initiatives, etc.
    - How does this differ between specific potentially risky technologies?
    - What seem to be the main drivers of these attitudes?
    - How much do these attitudes seem to influence China's policies on these matters?
    - What do the other major influences on those policies seem to be?

[24] This subtopic is partly inspired by questions from the post Research questions that could have a big social impact, organised by discipline.

- Community building, field building, and/or values spreading in China
    - What longtermism-relevant communities and fields already exist in China? How large and influential are they? What is their level of awareness of and connection to other longtermism-relevant communities and fields (especially the effective altruist longtermist movement)?
    - What longtermism-aligned values and attitudes are already prevalent in China? How amenable might various groups in China be to longtermism-aligned values and attitudes?
    - What is the importance, tractability, neglectedness, and downside risks of community building, field building, and/or values spreading in China, both in general and in relation to specific possible actions?
    - What would the best actions and framings for doing or supporting such work be? Which actors have a comparative advantage for these actions? Which actors should steer clear of these activities?
- Persistence or lack of persistence of certain Chinese institutions, norms, etc.[25]
    - Why have certain aspects of Chinese civilisation been so long-lasting? Are there any lessons we can draw from this about what makes for highly resilient institutions, cultures, or schools of thought?
    - What lessons can we draw from aspects of Chinese civilization that *didn't* last as long, or that lasted for a long time but eventually ended or changed?
        - It might be best to focus on aspects of Chinese civilization that seem in many ways similar to other aspects that were more persistent, as that reduces the number of variables that might explain the difference in persistence.
    - Why has Mohism almost died out in China, relative to other schools of thought?
        - What lessons can be drawn from this, for example in relation to the potential spread and persistence of schools of thought and communities related to utilitarianism or effective altruism, in China or elsewhere?

Please consider taking this brief survey to give feedback on this research agenda.

[25] This subtopic is adapted from questions from the post Research questions that could have a big social impact, organised by discipline. See also this research idea, which is also about persistence of institutions, norms, etc., but without necessarily focusing on China specifically.

Acknowledgements

This research agenda is a project of Rethink Priorities. It was written by Michael Aird. Thanks to Guive Assadi, Tobias Baumann, Janique Behman, David Rhys Bernard, Vicky Clayton, Shaun Ee, Daniel Eth, Juan Gil, Kit Harris, Hauke Hillebrandt, Peter Hurford, Alex Lintz, Darius Meissner, Konrad Seifert, Rumtin Sepasspour, Maxime Stauffer, Yip Fai Tse, and Linch Zhang for helpful comments and/or discussions. If you like our work, please consider subscribing to our newsletter. You can see more of our work here.