This doc is now superseded by the following public post: Research project idea: Intermediate goals for nuclear risk reduction. There's no longer any reason to read the doc version.

Research project idea: Intermediate goals for nuclear risk reduction

This post is part of a series of rough posts on nuclear risk research ideas, which I drafted in 2021 but only now got around to posting. ***I strongly recommend that, before you read this post, you read the series' summary & introduction post for context, caveats, and to see the list of other ideas.*** *I'm grateful to Will Aldred for help with this series.*

Some tentative bottom-line views about this project idea

Importance	**Tractability**	**Neglectedness**	***Outsourceability***
High	Medium	Medium/High	*Medium/Low*

What is this idea? How could it be tackled?

By an intermediate goal, I mean a goal that (1) is more specific and directly actionable than a goal like "reduce nuclear risk", (2) is of interest because advancing it might be one way to advance a higher-level goal like that, but (3) is *less* specific and directly actionable than a particular intervention (e.g., "advocate for the US and Russia to renew the INF Treaty").[1]

One version of this project would be copying and "finishing" a new and improved version of my not-properly-finished post "Intermediate goals for reducing risks from nuclear weapons: A shallow review". I recommend reading the Summary and Introduction of that doc to get a clearer sense of what that might look like and why it might matter. If you're interested in doing that version of this project, please reach out to me and we can discuss whether and how to proceed with that.

A more ambitious version of this project could in various ways go beyond what I was drafting. Ways of going beyond what I was drafting include:

1. Providing info on additional potential intermediate goals
2. Finding better or complementary ways to organize the goals

[1] I adopted the term "intermediate goal" from Muehlhauser (2020, 2021), who doesn't provide a definition but does give examples that illustrate the concept. The definition proposed here is my own.

Karnofsky (2022) also discusses a similar concept:

> 'How much should one value "transformative AI is first developed in country A" vs. "transformative AI is first developed in country B", or "transformative AI is first developed by company A vs. company B", or "transformative AI is developed 5 years sooner/later than it would have been otherwise?"
>
> If we were ready to make a bet on any particular intermediate outcome in this category being significantly net positive for the expected value of the long-run future, this could unlock a major push toward making that outcome more likely. I'd guess that many of these sorts of "intermediate outcomes" are such that one could spend billions of dollars productively toward increasing the odds of achieving them, but first one would want to feel that doing so was at least a somewhat robustly good bet.'

3. Providing more information about what some of the goal are, how easy it'd be to achieve them, what resources are most needed to achieve them, what effects achieving them may have on things other than nuclear risk, why they might decrease/increase/not substantially affect nuclear risk, and/or what interventions or organizations could be supported to help achieve them
4. Providing *better* (as opposed to "more") information and "bottom-line beliefs" on those matters than my draft does (i.e., info/beliefs that are more accurate, more focused on the most important points, and less misleading)
    - For example, someone could do or organize red-teaming of the post as a whole or its more important claims.
5. Changing the post or writing one or more new posts in such a way that it's easier and more likely for decision-makers to (correctly) use the post when making relevant decisions (e.g., making it easier for a decision-maker to find the info that's most relevant to them, or making other versions of the post that are tailored to particular target audiences)
    - For example, cutting out intermediate goals that seem low priority, providing more or less detail on each goal, providing more concrete intervention ideas.
6. Disseminating insights from the post to relevant decision-makers to increase the chance they act on them
7. Discovering and disseminating what various relevant actors/groups believe about these goals, to create common knowledge and aid in coordination[2]

A given project could do anywhere from just one to all six of those six things.

Each of those things could be done with a focus on just one potential goal, just a handful of potential goals, or a large number of potential goals. For example, a researcher do a deep dive into one of the possible intermediate goals to gather much more info, correct errors or misleading implications, write up their findings in a way tailored to whichever decision-makers are most relevant to the goal (e.g., EA community members making career decisions vs EA funders vs non-EA nuclear risk advocates vs US policymakers), and reach out to those decision-makers to discuss their findings. Or a researcher could spend 2-10 hours each on expanding and improving the info on >20 of the potential goals.

Specific actions that could be taken to tackle this project include:

- Reading more existing research, discussions, or opinions about the potential intermediate goals
    - For many goals, a *lot* has already been written, in some cases stretching over many decades, and I barely scratched the surface
- Quantitatively estimating the tractability or likely consequences of progress towards one or more potential intermediate goals
    - This could be done via making or soliciting forecasts, Fermi estimates, or more careful models
- Conducting (perhaps brief) impact assessments of previous efforts related to one or more of the intermediate goals (see also Research project idea: Impact assessment of nuclear-risk-related orgs, programmes, movements, etc.)

[2] Information about what various actors believe can prevent issues like some actors charging ahead due to not being aware that other actors see a given goal as net-negative, or conversely holding back from pursuing a given goal due to unfounded worries that other actors might see that goal as net-negative.

- Expert elicitation on the above points, via interviews, surveys, or convening workshops
    - This could include very open-ended questions like “What intermediate goals do you think people should focus on supporting”, somewhat open-ended questions like “Do you have any thoughts on the tractability or likely consequences of pursuing [specific intermediate goal] or which interventions would be best for that purpose?”, and/or rating scale questions
    - I’ve been involved in designing a similar survey focused on a different cause area and would be happy to provide advice, templates, etc.
- Creating “maps” or causal diagrams[3] of how various high-level goals connect to lower-level goals and then to specific intervention/policy ideas, to make it easier to understand what various goals being (very) net-positive or (very) net-negative would imply about what interventions to pursue/avoid

Why might this research be useful?

We lack clarity on what intermediate goals for nuclear risk reduction should be pursued and prioritized. Many goals that are commonly discussed may be intractable, might have little effect on nuclear risk even if achieved (e.g., because the goal would mainly just reduce the odds of a relatively low-stakes scenario), might *increase* nuclear risk if achieved (e.g., due to undermining deterrence), or might cause substantial harms unrelated to nuclear risk (e.g., increasing the odds of major non-nuclear armed conflict). Additionally, I’m aware of various goals that are rarely discussed but that seem to me plausibly worth prioritizing, and I expect more such potential goals could be discovered or thought of.

This seems to me like one of the key bottlenecks to our ability to reduce nuclear risk. This seems especially true given that the EA community has a large amount of money but relatively few committed members (especially those who have *and are known to have* various key skills, connections, etc.); under those conditions, we’d ideally just know what we want and be able to fund non-EAs to work towards that.

See also Muehlhauser (2020, 2021) for similar thoughts in relation to AI governance.

What sort of person might be a good fit for this?

This project idea is very broad and could be taken in many directions, so I think many people could work out and execute some version of it that’s well-aligned with their skills and interests. The project could also range from very deep and extensive research to taking relatively “simple” and “obvious” actions to improve the post I already wrote, so for any of a wide range of skill- and seniority-levels there’d be *some* version of this project that would fit well.

Some relevant previous work

- My Shallow review of approaches to reducing risks from nuclear weapons

[3] See also Causal diagrams of the paths to existential catastrophe.

- My Intermediate goals for reducing risks from nuclear weapons: A shallow review (part 1/4)

Should we try to convince/fund non-EAs to do this work?

I think deeper research or distillation of research on many of the specific intermediate goals would suit the skills and interests of many non-EAs, and is in fact similar to what many non-EAs are already doing. It might be worth trying to convince/fund non-EAs to do that work with a focus on the goals that seem most promising or where uncertainty is largest.

But I think it would be important to have an EA vet and extend the outputs of such work, such as by considering additional possible downside risks or considering in more detail whether and how the goal may reduce the odds or severity of especially worrisome nuclear conflict scenarios. And I think it would be important to have an EA to synthesize these various outputs into bottom-line views on what this suggests about how much to prioritize nuclear risk reduction and what to prioritize within that area. This could all happen after and separately from the non-EA research, or via an EA being part of the research team working on these outputs, or via EAs reviewing and giving feedback on the work.

It also seems very feasible to contract non-EAs to handle various tasks related to convening a workshop or designing, administering, and analyzing results from a survey, either (a) after an EA provides some of the content for these things and a clear explanation of the intended outcomes or (b) with the EA staying involved throughout this process.