Open questions in chapter 6 of *The Precipice*

Concrete questions

- What are the most important existential risks and risk/security factors to focus on? (Based on frameworks that are suggested in the chapter)
- What are ways to work on those risks & risk factors?

Measurement challenges

- How to apply ITN in practice?
    - How can the importance of different x-risks and x-risk factors be measured? [**not addressed in the chapter**]
        - How to estimate causal effects when the research subject has little or no empirical precedent?
        - How to integrate the correlation between risks & between risk factors in the estimate of their overall importance?
        - Anatomy of an x-risk as a heuristic? -> also does not offer concrete guidelines for assessing and ranking different potential risks and risk factors (how do you estimate p_origin, p_scaling, and p_endgame?)
    - How can the tractability of different x-risks and x-risk factors be measured? [**not addressed in the chapter**]
        - How to identify promising interventions, and how to measure their (expected and actual) impact?
        - How to integrate the correlation between risks & between risk factors when estimating the effects of specific interventions?
    - How can the neglectedness of different x-risks and x-risk factors be measured?
        - Claim: Neglectedness is measured by the amount of resources devoted to a risk/intervention/cause [**postulated in the chapter, no further argumentation**]
        - Challenge: Is it reasonable to compare different efforts at reducing risks by expressing them all in monetary values? Shouldn't the effectiveness of different efforts be factored in somehow (e.g., two projects to increase pandemic preparedness may each cost USD 1 million but have substantially different impacts; in the current definition of neglectedness, they would both reduce the neglectedness of work on pandemic preparedness by the same amount)?
- How should the ideas of soon, sudden and sharp factor into assessments? [**not addressed in the chapter**]
    - How to assess whether a given risk is more soon, sudden or sharp than another/all others?
    - How to incorporate different levels of soonness, suddenness, and sharpness into an assessment? How much impact should soonness, suddenness, and sharpness have on prioritization decisions?

    - Do these concepts make thinking about x-risk work more muddy/vague?
- How can we discover risks that are currently not even on our radar? What are the implications of the possibility of such risks for prioritization decisions and strategy-setting? [**not addressed in the chapter, but discussed elsewhere**]

Implementation challenges/questions

- How much to invest into improving risk estimation accuracy vs. directly investing into tackling risks? [**not addressed in the chapter**]
- How much to invest into figuring out effective strategies for tackling risks and risk factors vs. spending on actually implementing those strategies? [**not addressed in the chapter**]
    - Claim from book: “When a set of risks have equal tractability, or when we have no idea which is more tractable, the ideal global portfolio allocates resources to each risk in proportion to its contribution to total risk” (p. 182)
    - Possible challenge: maybe the ideal portfolio in that case should allocate substantial amounts of resources to the meta question of "how do we increase tractability of the risks in question? how do we get better at understanding causal links between possible interventions and risk reduction?"
- How to bring different prioritization frameworks together when deciding which causes and interventions to focus on? [**not addressed in the chapter**]
    - E.g. if risk A “scores” higher than risk B on ITN, but risk B is more sudden, soon, and/or sharp than risk A - how should that influence prioritization?
    - E.g.: How can the idea that risks are often correlated be integrated into an ITN-assessment of each individual risk (or of each risk factor)?
    - E.g.: How can work on the possible origin, scaling, and endgame scenarios for a given risk be integrated in an ITN-assessment of that risk?
- How to coordinate individuals and groups working on x-risks? [**not addressed in the chapter**]
- How do we rebalance our portfolio in the face of a changing risk landscape? [**not addressed in the chapter**]

Critical questions (questioning claims and assumptions in the chapter)

- Should prioritization between different existential risks and risk factors really be made based on the ITN-framework?
    - Claim (more or less implicit): Estimating I, T and N even roughly is better than not trying to measure them at all. Making the ITN-calculation with low confidence and high margins of error still adds valuable (and action-relevant) information. [**hardly addressed in the chapter, but discussed extensively elsewhere**]
    - Possible challenge (inspired by Nassim Taleb’s *The Black Swan*): If levels of uncertainty are high, maybe we should rely on a different framework that doesn't depend as much on predictive capacity. -> diversification, focus on resilience
    - Existing critiques of ITN: here and here
- Should prioritization between different existential risks and risk factors really depend only on the contribution to total existential risk? [**not addressed in the chapter, but probably (?) elsewhere**]

    - Possible challenge: Prioritization should not focus only on the contribution of an event to total existential risk; non-existential consequences with high importance in non-longtermist value systems should also be considered.
- Should the prioritization focus (of EA, of individual researchers, etc) really be on marginal impact? [**not addressed in the chapter, but elsewhere**]
    - Claim (e.g. p. 181): "Where can an additional bundle of resources (such as time or money) most reduce total risk?"
    - Possible challenge: The focus should instead be on the end goals, and actions/interventions/projects should be evaluated based on their absolute importance for reaching that end goal, not based on conjectures about what other people/groups will do.
    - Possible challenge: The focus should not be on the added value of an additional unit/bundle of resources, but instead the focus should be on re-balancing existing resources.
    - Possible challenge: Are there really diminishing returns to work on different existential risks and risk factors?
- How much should we focus on correlations between risks? [**hardly addressed in the chapter**]
    - How much do correlations between risks matter for assessing their importance?
    - Can we estimate correlations between risks reliably?
    - Is it justified to “expect some positive correlation between most pairs of risks" (p. 174)?
    - Concrete claims: p. 175 and Appendix D
    - Is the following “robustness check” a useful/valuable/insightful procedure for improving prioritization decisions?: start assuming risks are independent, then assume they are fully correlated, then fully anti-correlated, then compare and see how model changes (as advocated on p. XXX)
- Are existential risk factors really a useful framework for thinking about how to reduce total existential risk?
    - Is it true that there are some events, developments and conditions/circumstances that are unlikely to result in existential catastrophe themselves, but which may make existential catastrophe as a result of some other risk more likely? [**hardly addressed in the chapter**]
    - Can we identify and prioritize between these risk factors reliably/robustly? [**not/hardly addressed in the chapter**]
        - Is it really easy to identify "stressors for humanity or for our ability to make good decisions" (p. 179)?
        - On prioritization: see questions about measuring ITN above
    - Are there things we can do to tackle these risk factors effectively? [**not addressed in the chapter**]
- Should individuals and groups really place great importance on individual fit and leverage when deciding which existential risk or risk factor to focus on? [**hardly addressed in the chapter, but discussed extensively elsewhere**]

    - Possible challenge: maybe different existential risks are so different in importance that it would make sense for individuals or groups to take a long time re-skilling.
    - Possible challenge: Work on different existential risks doesn't differ all that much by personal fit. There are different ways of addressing an existential risk, and most people will find a way that fits them for each risk.
- What are the claims about concrete existential risks and risk factors in the book based on? How robust are they? [**all of these are not or only poorly substantiated in the chapter, but some or all may be addressed elsewhere and/or have quite high intuitive appeal/plausibility**]
    - Importance of Great power war as an existential risk factor
        - "For example, it seems that the bulk of the existential risk last century was driven by the threat of great power war." (p. 149).
        - "Consider your own estimate of how much existential risk there is over the next hundred years. How much of this would disappear if you knew that the great powers would not go to war with each other over that time? It is impossible to be precise, but I'd estimate an appreciable fraction would disappear—something like a tenth of the existential risk over that time. Since I think the existential risk over the next hundred years is about one in six, I am estimating that great power war effectively poses more than a percentage point of existential risk over the next century. This makes it a larger contributor to total existential risk than most of the specific risks we have examined." (p. 149)
    - There are probably only a few important risk factors (p. 180)
    - "These [stressors for humanity or our ability to make good decisions] include global economic stagnation, environmental collapse and breakdown in the international order.24 Indeed even the threat of such things may constitute an existential risk factor, as a mere possibility can create actual global discord or panic." (p. 151)
    - "Many risks that threaten (non-existential) global catastrophe also act as existential risk factors, since humanity may be more vulnerable following a global catastrophe. […] nuclear winter or climate change [...] could easily cause major catastrophes that leave us more vulnerable to other existential risks." (pp. 151-152)
    - "Examples [of existential security factors] include strong institutions for avoiding existential risk, improvements in civilisational virtues or peace between great powers." (p. 152)
    - "Many of the things we commonly think of as social goods [...] such as education, peace or prosperity" (p. 152) are probably existential security factors.
- What are the claims about concrete interventions/strategies in the book based on? How robust are they?
    - Targeted interventions are to be prioritized over broad ones because of neglectedness

- Because things in the long future are hard to predict, the focus should be "on knowledge and capacity building, over direct work." (p. 185)
- Individuals and smaller groups are likely to do best by "putting [their] efforts into a single risk" (p. 182).