HOW TO COMPARE BROAD AND TARGETED ATTEMPTS TO SHAPE THE FAR FUTURE

Nick Beckstead

Research Fellow, Future of Humanity Institute

Board of Trustees, Centre for Effective Altruism

Background assumptions of the talk

- This talk is about how to evaluate options if you accept the view that shaping the far future is overwhelmingly important.
- I'm not going to argue that shaping the far future is overwhelmingly important today. If you want to know what I think about that, you can read my PhD thesis from my website.

Rough outline of why the far future is overwhelmingly important

- The future could be big.
- A big future would be overwhelmingly important if we could affect it.
- We can affect the big future.
- This is enough to make far future considerations dominate altruistic choices.

How could you change the far future?

- There's a spectrum here from very broad to very targeted
- Very **broad** end of the spectrum: "We'll just empower people today as much as possible. This will make it more likely that, whatever challenges we face later, people will handle them better. And that's our best shot at putting humanity on a positive trajectory for the long run."

How could you change the far future?

- Very **targeted** end of the spectrum: “We should identify specific risks or scenarios that might play a pivotal role in the future of humanity, and try to make sure that those risks and scenarios are managed as well as possible. This is our best shot at positively influencing humanity’s trajectory in the far future.”
  - Main example of this: Eliezer Yudkowsky
- And there is a lot of middle ground. You might think that certain key factors are highly important for shaping the far future, and that we should emphasize those. These might include things like:
  - Coordination: How well-coordinated people are
  - Capability: How capable individuals are of achieving their goals
  - Motives: How well-motivated people are
  - Information: To what extent people have access to information
- That’s a list I like anyway.

Examples of targeted proposals

- Highly targeted proposals
  - Do technical research which will help build a Friendly AI
  - Advocate for nuclear disarmament to prevent a nuclear war
  - Reduce carbon emissions so that climate change is a smaller problem
- Some moderately targeted proposals
  - Tell people about the importance of shaping the far future, so that they make better decisions about that
  - Tell people about the importance of helping animals, so that they make better decisions about animals in the distant future
  - Do research on risks and opportunities from future technologies so that, in the coming decades, people make better choices about future technologies

Examples of broad proposals

- Some very broad proposals
  - Help make computers faster so that people everywhere can work more efficiently
  - Change intellectual property law so that technological innovation can happen more quickly
  - Advocate for open borders so that people from poorly governed countries can move to better governed countries and be more productive
  - Go work for Wikipedia to help improve the site's overall functionality

Examples of broad proposals

- Some pretty broad proposals
  - Meta-research: improve incentives and norms in academic work to better advance human knowledge
  - Education
  - Advocate for political party X to make future people have values more like political party X
  - Improve Google's search engine so that people have better access to information
  - Advocate for effective altruism so that people care more about doing good

What I think

- I’m highly uncertain about where on the spectrum we should be
- I think a lot of the broad stuff, including technological progress, is positive, very unclear on the effect size
- I think a lot of people doing stuff that is good by broad standards aren’t aiming at the far future at all, and their efforts could well be competitive with people doing highly targeted stuff. But my views on this are not stable.

Arguments I'll make

- I'm going to say mostly things in favor of broad approaches because I think people who care about the far future underemphasize them
- I'm going to try to make many rough, independent, weak arguments, rather than trying to craft one relatively strong argument
- I'm not going to go into great detail on any of these arguments, so there are important objections and replies I am not considering

ARGUMENTS FOR BROAD APPROACHES

Broad approaches are more conventional

- This speaks in favor of assuming that broad is better by default
- But it's a bit of a weak point since we have very little sense of how much this is driven by it being highly unconventional to think the far future is overwhelmingly important
- Still, I think most conventional, smart people would, by default, be more enthusiastic about enhancing education than worrying about specific future scenarios as a method of making the far future go better

Moral/coordination benefits from faster growth

- *The Moral Consequences of Economic Growth* by Benjamin Friedman
  - "Economic growth—meaning arising standard of living for the clear majority of citizens—more often than not fosters greater opportunity, tolerance of diversity, social mobility, commitment to fairness, and dedication to democracy."
  - And he argues that stagnation pushes in the other direction
  - Detailed historical arguments for this
  - Some theoretical arguments—I won't get into it but I find it plausible
- I have the general sense that countries are less likely to get into wars when things are going well economically. Some historians believe that stagnation was a significant factor in WWII.

Broad approaches and past challenges

- It seems that more effective broad approaches would have, in the past, resulted in better outcomes when civilization faced new challenges
- Examples that I think support this:
  - WWI and mustard gas
  - WWII and nuclear weapons (I'll highlight this one)
  - Cold War

Dissecting an unprecedented challenge: nuclear weapons in WWII

- What was the risk?
  - There was a concern about whether using the weapons would ignite the atmosphere and induce a global catastrophe
  - Possibly one country could have used a nuclear advantage to achieve world domination, for good or ill
  - Possibly the weapons could have been used on a larger scale, with massive short-term human consequences and major ripple effects

Dissecting an unprecedented challenge: nuclear weapons in WWII

- What determines whether this goes well?
  - When the transition is triggered…
    - Information: How well do the key actors understand the threat?
    - Coordination: How much trust and cooperation is there between countries?
    - Capability: How capable are the key actors?
    - Motives: To what extent do the key actors have the right motives?
- What was the trigger for the risk?
  - Nuclear physics reaching a certain level of understanding

You can now think about different interventions…

- If you speed up general technological progress…
  - Information: no clear effect
  - Cooperation: arguably less likely to be in a war if people were doing better economically
  - Capability of key actors: no clear effect
  - Motives of key actors: arguably better
- Seems good!

An argument I think is mistaken

- Note that this observation conflicts with an argument I have heard in conversation. The argument says:
  - Most existential risk comes from dangerous future technology.
  - If we have a higher rate of general technological growth, we'll get dangerous future technology sooner.
  - Dangerous future technology is very dangerous!
  - Therefore, it would be bad to have a higher rate of general technological growth.

An argument I think is mistaken

- This argument is misleading and highly incomplete for a few reasons:
  - It doesn't consider the fact that a higher rate of technological progress also speeds up factors that make us more prepared when we meet transition risks
  - It doesn't consider the fact that higher rate of progress can reduce state risk
  - It doesn't consider the fact that a higher rate of progress may make a technological stagnation less likely, which would also be good. This is a consideration I am less familiar with, but would like to investigate more closely in the future.

Wrapping up on nukes…

- I generally think that people doing better broad work would have been helpful.
  - Better educated people doing the work would have been good
  - People who were more thoughtful would have been good
  - People not in an economic stagnation might have meant no war, which would have been good
  - Better access to information in general would have been good (I imagine it might have been extremely helpful if the people on this project, e.g., had access to something like Wikipedia at the time)
- I have similar views on previous unprecedented challenges humanity has faced

A further note

- If you look at these areas (economic growth and technological progress, access to information, individual capability, social coordination, motives) a lot of everyday good works contribute
- An implication of this is that a lot of everyday good works are good from a broad perspective, even though hardly anyone thinks explicitly in terms of far future standards
- This puts limits on how special we think we can be by focusing on far future standards

Broad approaches benefit from economies of scale

- When we make search algorithms more efficient, a very significant portion of humanity can be empowered. If you provide information that future people can use for a specific scenario, a much smaller number of people facing pivotal challenges will be empowered (though they will be empowered more).
- Broad approaches tend to get more improvement in (impact per person * # of people affected) than targeted ones

Broad approaches and future needs

- Make it easier to tell how big of a problem you're dealing with, and what the "room for more resources" is
  - Comparatively easy to tell whether cancer research is currently underfunded
  - Comparatively hard to tell whether cancer research will be underfunded in 40 years
  - Comparatively easy to tell whether AI researchers currently know enough about AI safety, were AI to come now
  - Comparatively hard to tell whether, in 30 years time, AI researchers will know enough about AI safety

Should people solve the problem later?

- In some ways, trying to help future people navigate specific challenges better is like trying to help people from a different country solve their specific challenges, and to do so without intimate knowledge of the situation, and without the ability to travel to their country or talk to anyone who has been there at all recently
- Sometimes, only we can work on the problem (this is true for climate change and people who will be alive in 100 years)
- It is less clearly true with risks from future technology

Broad approaches and evidence/ feedback

- With broad approaches, it is generally easier to tell whether you are getting closer to your goal and readjust your course
  - I can tell whether my search algorithm is getting better, whether people are publishing more replications, whether the burden of malaria is falling.
  - If you are wrong about what is needed in a specific future scenario, it is easy to waste your time
- Though broad approaches also depend on speculative claims about long-term effects, and you can't really get relevant feedback on that

IN FAVOR OF TARGETED APPROACHES…

Most people who care about the far future are into targeted attempts

- This is relevant
- It is somewhat unsurprising because it is more important to think about far future considerations if you want to go targeted than it is if you want to go broad
- The ideas have been fairly bundled together by their leading advocates (Bostrom, Yudkowsky)

Targeted issues may have high room for more funding/talent

- From a GiveWell blogpost reviewing how the top 100 foundations allocate their money:
  - “Mitigation/prevention of global catastrophic risks other than climate change. 2 foundations focus on nuclear nonproliferation, while one focuses on biological threats; the total giving for this category according to dollar allocation data is 0.1% of all giving dollars.” - See more at: http://blog.givewell.org/2012/05/08/what-large-scale-philanthropy-focuses-on-today/#sthash.YycfLyyV.dpuf”
  - I note that it is very unclear what the percentage should be

Different values → funding gaps?

- Few people think the far future is overwhelmingly important, so they may overlook some targeted stuff
  - But people look at it for other reasons. E.g., GCRs matter to governments and militaries for obvious reasons, and you get scientists who want to know all about asteroids because they find them fascinating

Broad attempts can be a mixed bag

- Broad approaches are more likely to enhance bad stuff as well as good stuff
  - Increasing people's general capabilities/information makes people more able to do things that would be dangerous, offsetting some of the benefits of increased capabilities/information
  - Improving coordination or motives may do this to a lesser extent

Some scenarios really do look predictably pivotal

- AI looks like a foreseeable source of changes in our future development trajectory

- GCRs stand out as an area where changes to our development trajectory look unusually foreseeable

If you get it right, you can win ridiculously hard

- Can give you opportunities for outsized impacts if you have the right model of the risk and you identify real ways of making the risk go better that won't be solved in the future
- It's a type of strategy that is very challenging to do right, but is potentially very promising if you do it right

Little systematic thought has gone into far future considerations

- There seems to have been very little careful thought about how we might shape the far future in a targeted way. But it would be extremely promising if we found ways of doing it. So there is an argument for looking for targeted attempts, and closely vetting options that look plausible.
  - However, it's also true that little systematic thought has gone into thinking about how we can best shape the far future in a broad way. And little systematic thought has even gone into how we can accomplish as much good as possible by conventional or near-term utilitarian standards.

KEY QUESTIONS FOR FURTHER INVESTIGATION

Key questions for further investigation

- Is there a common set of broad factors which, if we push on them, systematically lead to better futures? (My current list is: social coordination, individual capability, individual motives, and availability of information.)
- Does the future depend on how humanity handles a small number of challenges? Can we tell what they are right now? Can we tell what to do about them? Could further research illuminate these questions?

Key questions for further investigation

- In history, how often did big wins (and failures) come from people addressing challenges that humanity would face in the distant future? Do the big wins and failures have common features? How often did people try this?
- In history, how often did big wins (and failures) come from people improving humanity's ability to address future challenges in general? Do the big wins and failures have common features? How often did people try this?

Key questions for further investigation

- What is the current space of opportunities for money and talent aiming to solve specific challenges that humanity will face in the future? Where are the resource gaps? What looks tractable? GiveWell is doing research relevant to this question.
- What is the current space of opportunities for money and talent aiming to enhance humanity's ability solve general challenges that humanity will face in the future? Where are the resource gaps? What looks tractable? GiveWell is doing research relevant to this question by looking into GCRs.

TAKE-AWAYS

Take-aways

- There is an interesting question about where you want to be on the targeted vs. broad spectrum, and I think it is pretty unclear
- Lots of ordinary stuff done by people who aren't thinking about the far future at all may be valuable by far future standards
- Broad approaches (including general technological progress) look more robustly good, but some targeted approaches may lead to outsized returns if done properly
- There are many complicated questions, and putting it all together requires challenging big picture thinking. Studying targeted approaches stands out somewhat because it has the potential for outsized returns.