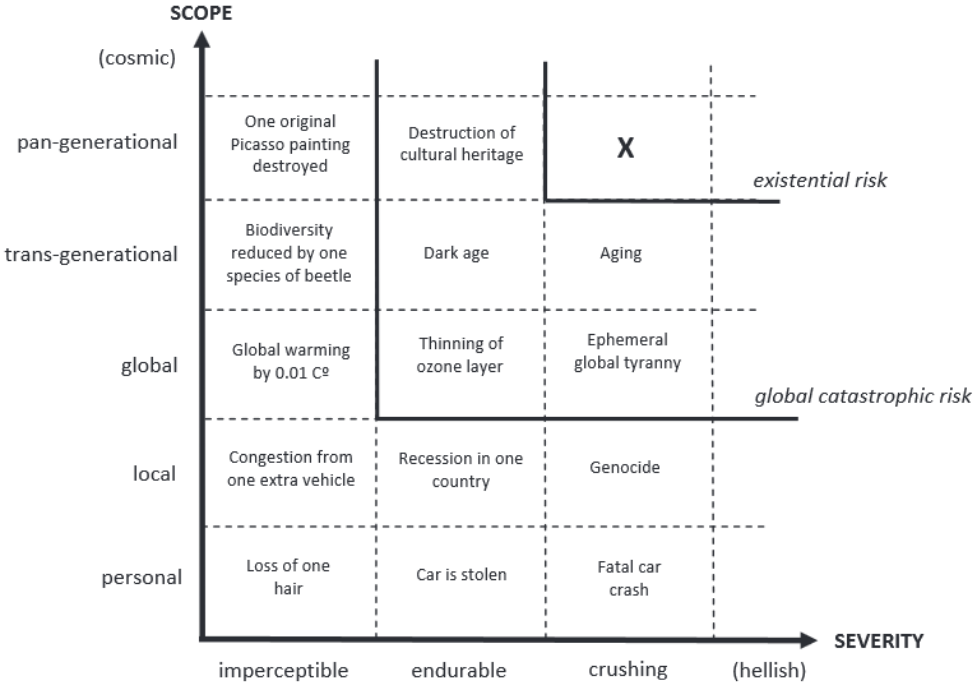
Week 3: Existential risk

Owen Cotton-Barratt

What kind of event or process could cause human extinction, or otherwise curtail our potential? How concerned should we be?

1. Qualitative risk categories

1. Catastrophes come in very varied levels of scope and severity. See diagram from Bostrom (2013):



   a. Could separate out scope into spatial and temporal components, but doesn't change the picture too much.
2. Generally we care more about things with larger scope and greater severity.
3. We will focus on top-right, existential risk.
   a. "An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development."
4. Are there any existential risks?
   a. Yes.
      i. Clearest case is threat of human extinction from very large natural disaster.
         1. Asteroid. Gamma ray bursts.
      ii. Maybe more threatening are anthropogenic risks.

1. Nuclear weapons gave for the first time potential to wipe ourselves out.

2. Natural risks

1. Asteroids
    a. A large enough asteroid impact could cause devastation.
    b. Primary mechanism for damage is throwing dust up into atmosphere. This blocks light, preventing plant growth (hence problems with food production).
2. Supervolcano
    a. Occasionally (~once in 70,000 years) get very large volcanic eruptions.
    b. Again, primary mechanism for damage is throwing dust up into atmosphere.
3. Pandemic
    a. Pandemic influenza appears largest threat.
        i. Some influenza pandemics infect most people.
        ii. Some strains of influenza have very high case-fatality rates.
            1. H5N1 has killed around two-thirds of people infected.
        iii. No in-principle reason these should not occur together.
4. Bounding natural risk
    a. Can use the historical record to give some bound on natural risk from all sources.

3. Anthropogenic risks

1. Note on probabilities
    a. When talking about these risks, probabilities are meant in the sense of 'what a reasonable observer would believe, given the evidence' rather than 'if we ran the world many times, the proportion of times something would happen'.
2. Nuclear war
    a. Could kill a large fraction of population directly.
    b. Risk of extinction again comes primarily from atmospheric effects ('nuclear winter').
3. Extreme climate change
    a. Climate change not projected to be bad enough that it could pose an existential risk, but there is a tail risk that it is significantly worse than expected.
4. Artificial pandemics
    a. Already possible to synthesise virus DNA
    b. The 1918 flu has its genome published.
    c. Future developments may allow engineering of even more deadly pathogens.
        i. Some of this is already occurring! Some benefits, but potentially large risks.

5. Global totalitarianism
    a. Enabled by appropriate technologies, we might enter a state which is persistent and destroys most of the value of the future without causing extinction.
6. Artificial intelligence
    a. Plausibly a very big determinant of the long-run direction of humanity.
    b. Some things we can say, but a lot we don't know.
    c. Will treat in more depth in Week 5.
7. Hard to give precise probabilities of anthropogenic risks, but many judgements put them between 1% and 50% over the coming century – anything in this range is much higher than bounds on natural risk.

4. Why care? A more formal argument

1. Assumption: future could be very large.
    a. Last week Toby Ord painted a picture of how it could get extremely large – spread over billions of years and billions of galaxies.
    b. Even a more modest assumption that the Earth could support 5 billion people for 1 million years (lifetime of typical mammalian species) would mean 50,000 billion lives of a century each.
2. Assumption: human lives are often significantly worth living.
    a. Quite uncontroversial!
3. Assumption: bringing someone with a life worth living into existence is a good thing to do.
    a. Many people have this intuition.
    b. More controversial. Many people have "person-affecting" intuitions, that actions are only good or bad to the extent that they affect people who already exist.
        i. But there are some issues with person-affecting views. Suggest we don't need to worry about long-term climate change, because any actions we take will change the collection of people who will be around.
    c. This is a deep subject – taking it as an assumption for now. See "Weighing Lives" by Broome for deeper discussion.
4. Assumption: creating good lives is broadly aggregative. The value of another good life doesn't depend too sensitively on how many already exist.
5. **Conclusion A**: The value of the future may be very much larger than the value of the present.
6. Assumption: We can affect the long-term future in (in expectation) non-negligible ways.
    a. Reducing natural risk of human extinction gives a particularly clear example.
        i. Not a claim that this is the best way to affect the future!

7. **Conclusion B**: Most of the value we can affect is in the long term.
8. Assumption: In futures with a lot of eventual social change, it is nigh-impossible to predict the direction of effects of actions today.
9. **Conclusion C**: "Maxipok" principle
    a. **Maxi**mise the **p**robability of an **OK** outcome.
        i. Where an OK outcome is any that avoids existential catastrophe.
        ii. This is a suggested heuristic for impersonal moral action.
10. References:
    a. "Existential Risk Prevention as Global Priority", Bostrom (2013)
    b. "On the Overwhelming Importance of Shaping the Far Future", Beckstead (2013, PhD thesis)
11. Other possible arguments
    a. The argument presented here is broadly consequentialist.
        i. Scope for someone to explore how well an analogous argument from the rights of future generations might work. Are the conclusions different?
    b. Even if we only care about present generation, seems extinction risks may be neglected.

5. Why care? A common sense argument

1. A common reason for caring about climate change is a sense of stewardship of something valuable. We are in a position to look after the planet; we ought to do so.
2. Humanity is also something precious. We have a duty to preserve it, and its future prospects.
    a. If the Romans had developed nuclear weapons and ended humanity in a war, with an outside perspective we'd say they screwed things up!
3. So why do we not already have cultural norms about this?
    a. Only relatively recently that we developed capabilities which could affect existential risk.
        i. Development of thermonuclear weapons in the 1950s.
            1. Much of the time since then was in the cold war.
        ii. Ability to start protecting against asteroid risk.

6. Open questions

- What could destroy us? From our perspective, what are the likelihoods of different risks?
- Under exactly which assumptions (ethics, discounting, etc.) are these risks a large concern?
- What do sustainable trajectories look like?
- What sort of institutions might avert the market and political failures