Week 8: Differential progress and information hazards

Owen Cotton-Barratt

The long-term trajectory of humanity is likely to be affected by the direction of progress more than its speed. How can we identify and affect this? When are we better not sharing information?

1. Question: when is progress good?

1. Common sense answer: progress seems generally great!
    a. World is much better today than centuries ago.
2. Long-term perspective
    a. Progress still essential, to attain large valuable futures
        i. Astronomical waste argument shows that speed of progress is extremely valuable.
        ii. See Ord on *Grand Futures* in Week 2 for more details.
    b. Existential risk argument implies best to pursue progress in the manner which will minimise total risk.
        i. See Week 3 on *Existential Risk*.
        ii. We'll look at some models for this effect.
3. Simple model: risk per unit time is a function of level of progress.
    a. Simple in modelling progress as one-dimensional.
    b. Presume risk tails off with sufficient progress.
    c. Best strategy is to go as fast as possible to minimise total exposure.
4. Models with multidimensional progress
    a. e.g. technology, coordination
    b. Get landscape of risk levels
        i. Total risk now depends not just on *speed*, but also on *trajectory* through this landscape.

2. Differential progress

1. Is it better to try to change speed or trajectory?
    a. Depends on:
        i. Which has greater effect on total risk?
        ii. Which we can affect more?
    b. I think 'trajectory' is plausibly best on both counts.
2. Sustainable trajectories
    a. Perspective where choosing a good trajectory is a sustainability issue.
3. Principle of differential progress

    a. Understand that we probably cannot change which technologies will eventually be developed, but we may be able to change the sequencing.
    b. Hence it is good to retard development of risky technologies, and accelerate progress which would help to avoid or mitigate these risks.
        i. Complication: some technologies may do both at once!
        ii. Typically there are better levers for accelerating than retarding.
    c. cf. 'differential technological development'
4. Aiming at safe versions of a technology
    a. From our current perspective, dangerous versions of some future technologies look almost the same direction as safe versions.
        i. Can try to abstract the *difference*, and push in that direction.
5. What directions are robustly good?
    a. Increasing/improving coordination
    b. Insight / wisdom
    c. Focus on safety

3. Information hazards

1. Differential progress suggests that not all technological progress is beneficial.
2. More generally, we can ask: when does dissemination of (true) information cause harm?
    a. General social norms towards truth and knowledge!
    b. But several acknowledged exceptions:
        i. To protect security (national or personal)
        ii. To preserve jury impartiality, or otherwise avoid bias
        iii. Preserve anonymity for patients, reviewers, voters, etc.
    c. A threat of harm from such dissemination is called an *information hazard*
        i. Sometimes the term is transferred to the information itself.
3. Many types of information hazard, e.g.
    a. Leak of passwords online
    b. Could disrupt helpful norms
    c. Spoilers
4. Main types with relevance to existential risk
    a. Developments of new technologies that are bad from a differential progress perspective
    b. Developments that could lead to new norms which are less stable or robust than old norms.
    c. Information that could give enemy actors more powerful tools
    d. Information that could shift attention in unhelpful directions
        i. e.g. perhaps premature discussion of risks makes people defensive
5. Note that information can take many forms
    a. Traditionally we'd think of data/ideas

    b. But even if something is in public domain, disseminating it further can have effects.
        i. So the act of pulling attention to it, and even the way in which it is communicated, can constitute an information hazard.
6. Major reference with much more detail on different types: *Information Hazards: a Typology of Potential Harms from Knowledge*, Bostrom (2011)
7. Some domains where we have well-established norms for dealing with information hazards. But other areas where we don't have these yet, and there is useful work to be done in building and refining them.

4. Open questions

- How should information hazards be treated?
    - By individuals, institutions, society
    - In particular, how should a research institute considering existential risks treat information hazards?
- Is it correct that we have more effect via trajectory than via speed?
- For major societal variables like economic growth, technological progress, population, what are the main routes for their effect on long-term trajectory?
- Rich vein of questions:
    - for many questions of policy today, which direction is in expectation better for the long-term?