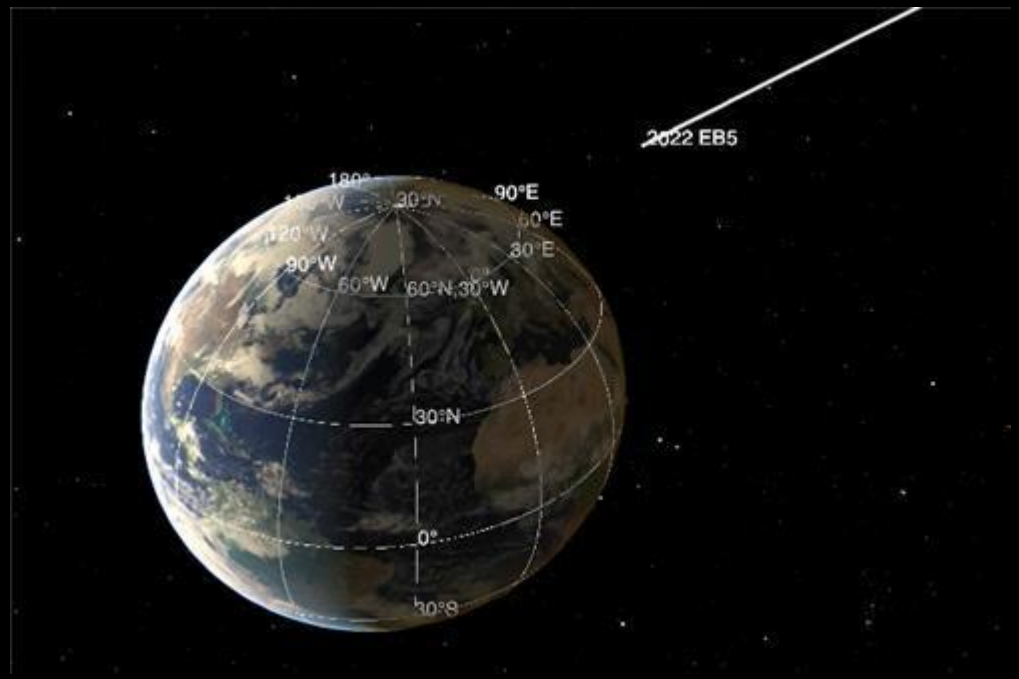
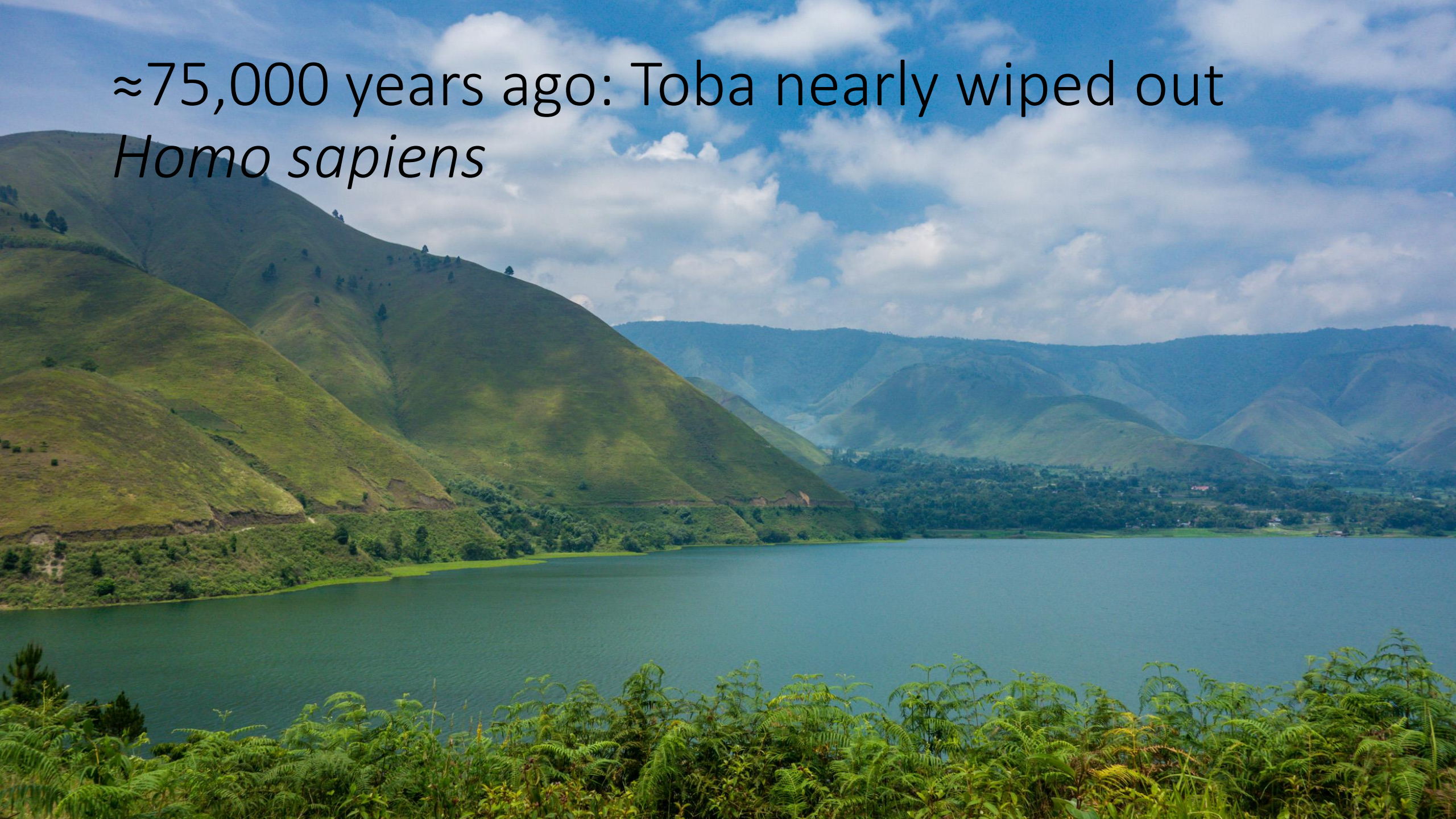
Global and existential risks

Anders Sandberg
Reuben College
Future of Humanity Institute
University of Oxford

March 11 2022: asteroid 2022 EB5 explodes north of Iceland



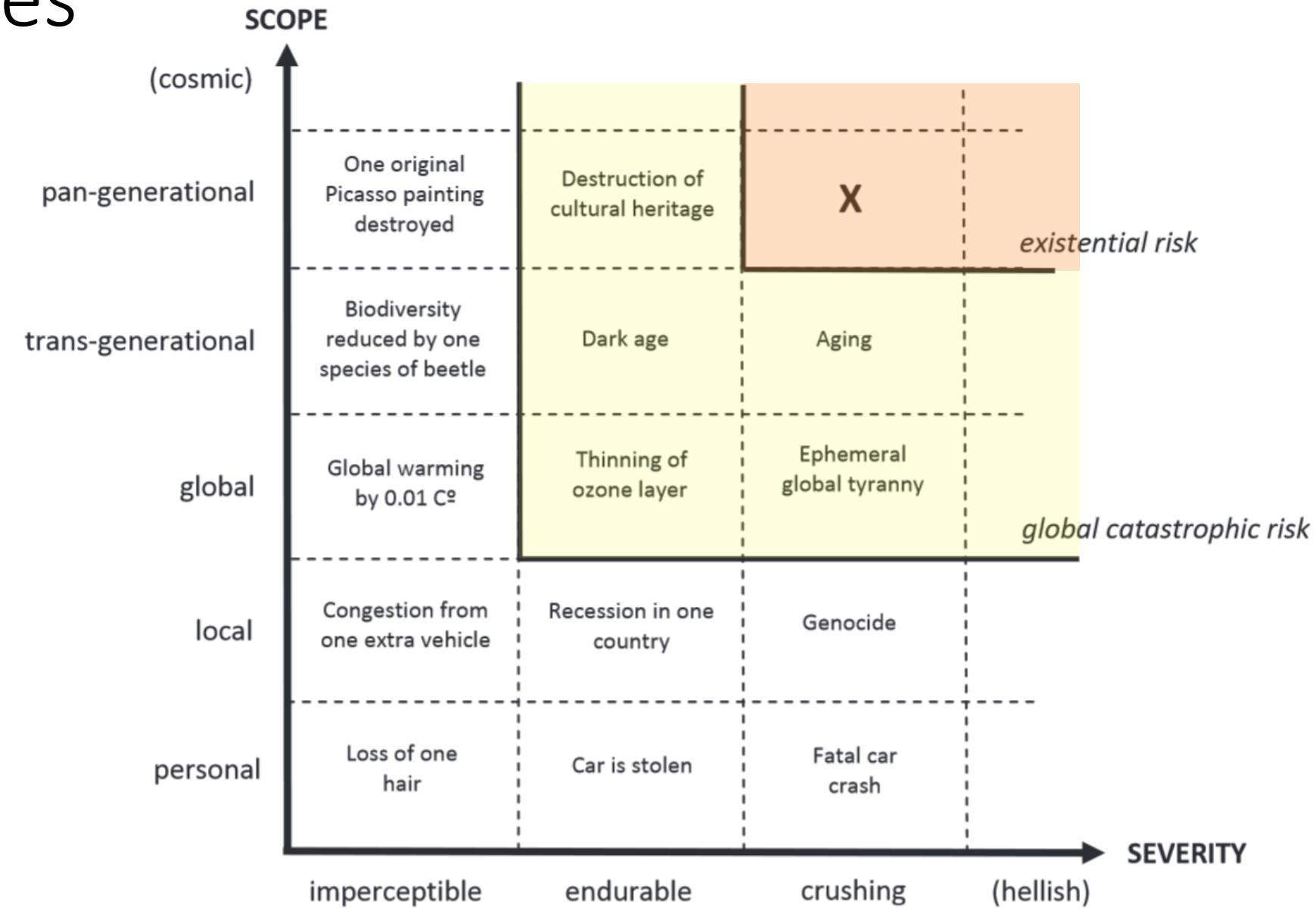
September 26 1983: Stanislav Petrov saves the world



≈75,000 years ago: Toba nearly wiped out *Homo sapiens*

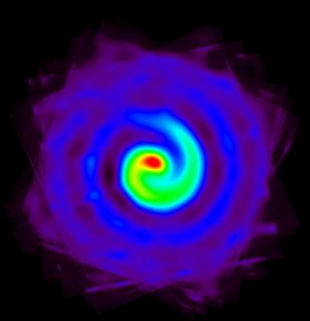
Risk categories

Risk impact is a combination of scope (how many affected) and severity (how much harm)

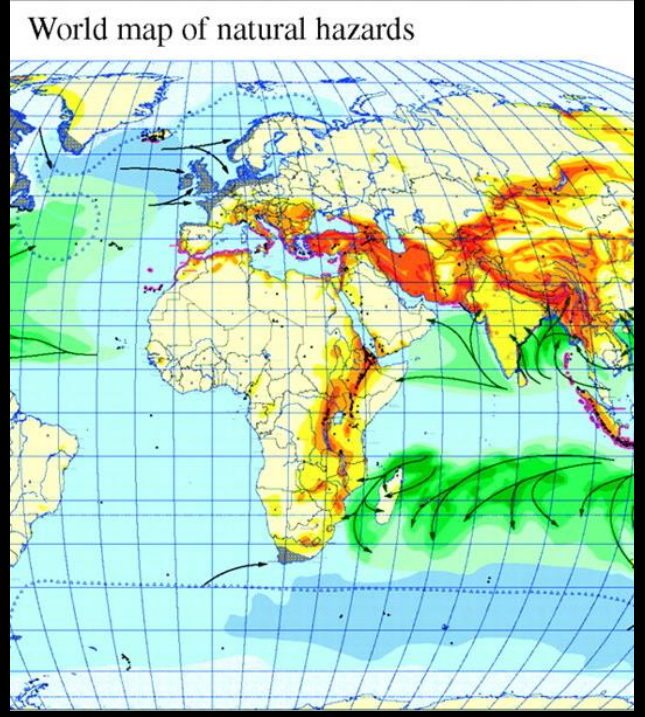


Badness of existential risk

- Loss of value
  - Loss of value in remaining lifespan of existing humans
  - Loss of value of all future lives that won't happen
  - Loss of value of all human culture and artifacts
  - No future culture and artifacts
  - No future value of the species itself
  - Loss of valuers
- Breaking the continuity of humanity
- Fair risk distribution
  - Do not discriminate against future people
  - Global risks fall very unequally
- Loss of options

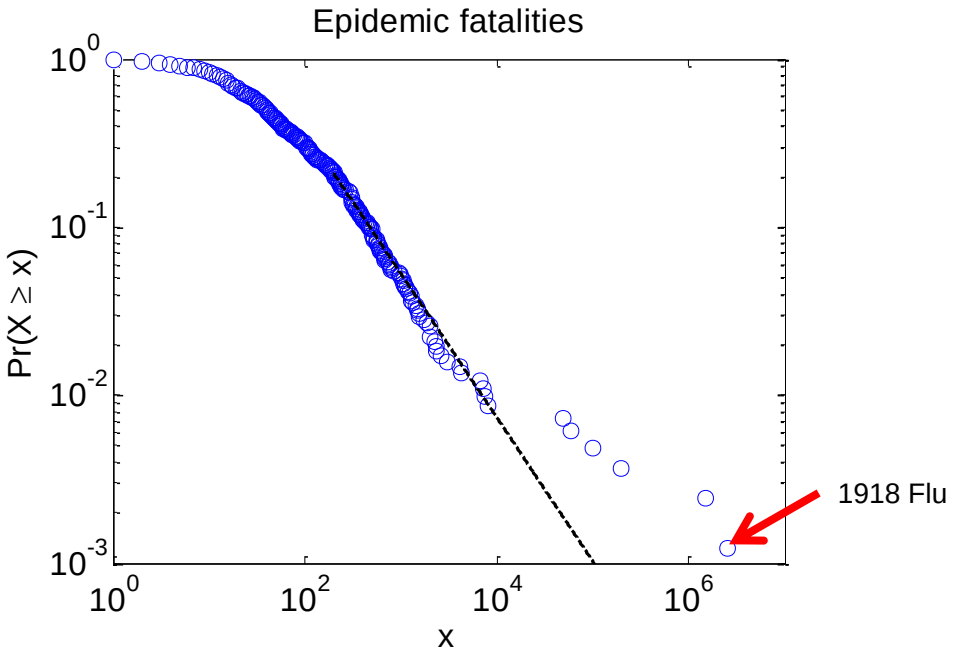


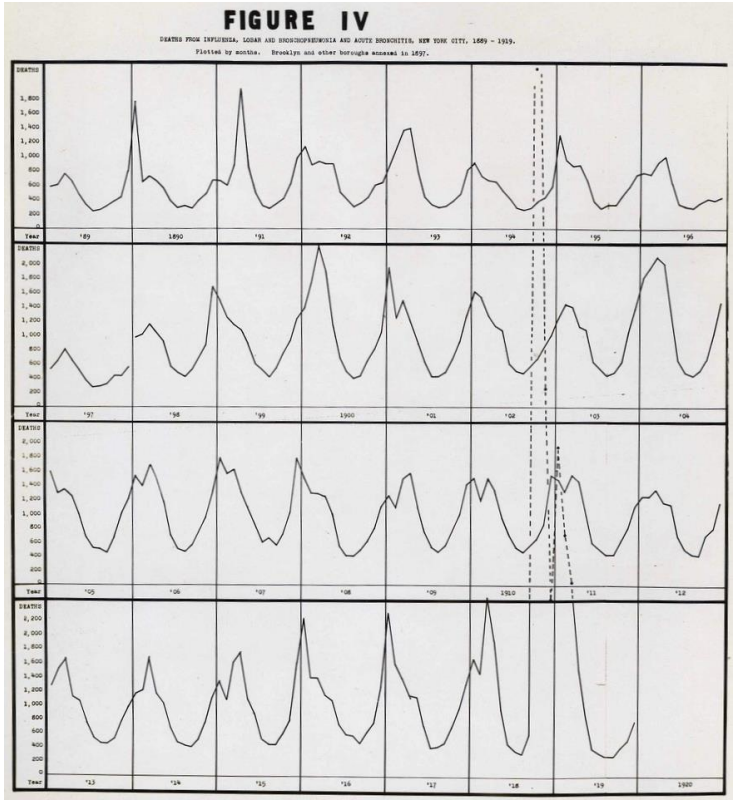
Astronomical risks

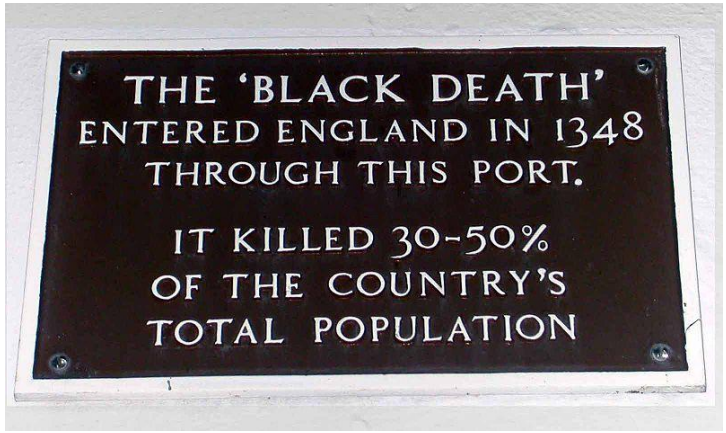


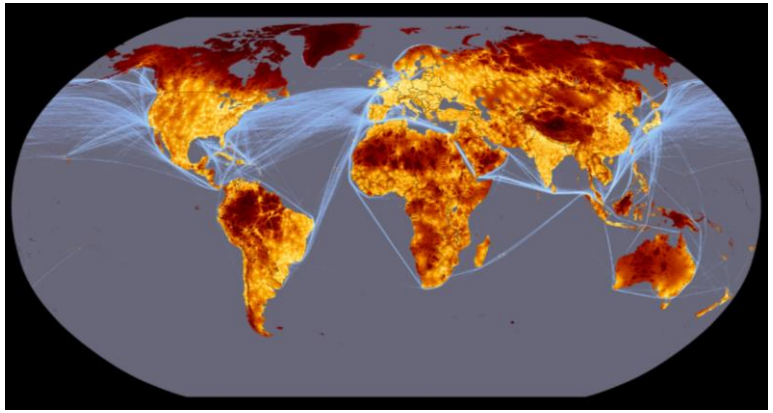
Geophysical risks

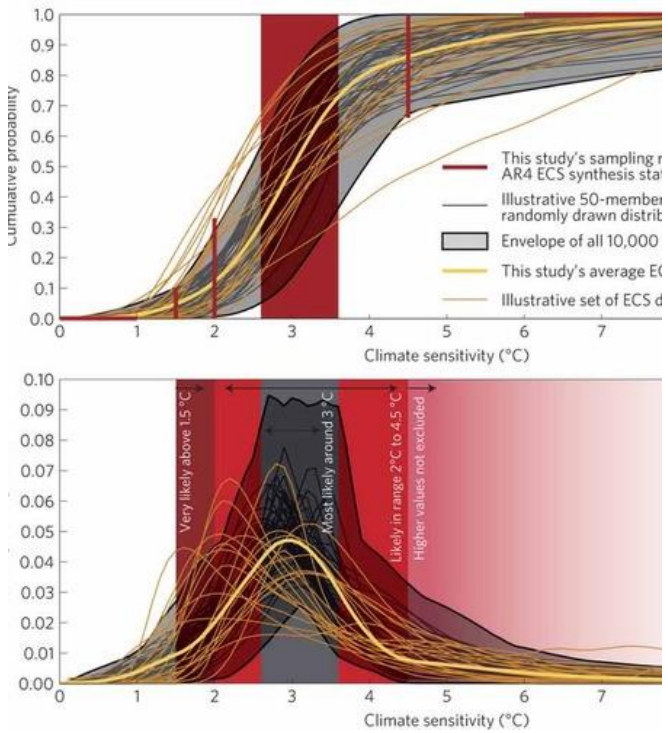
Natural pandemic risks



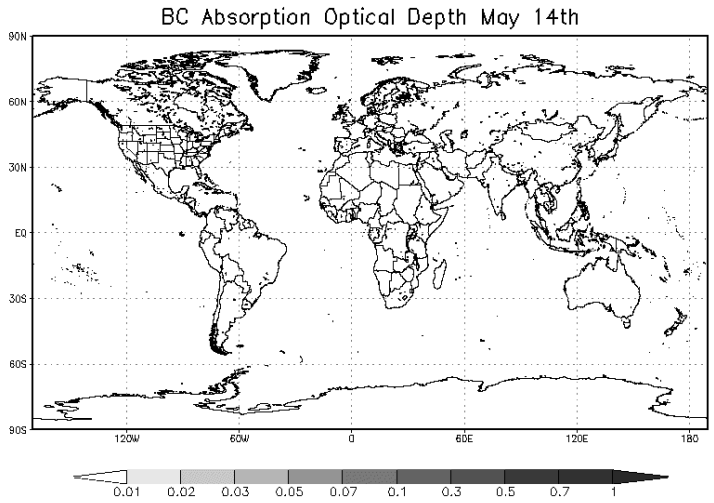








Anthropogenic risks



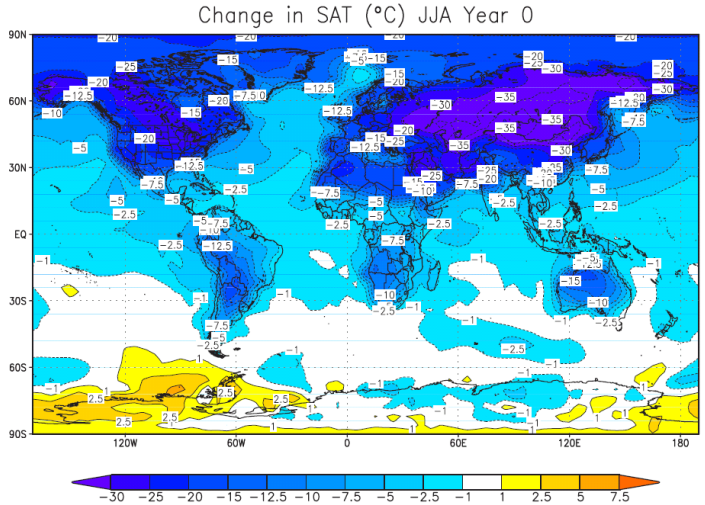
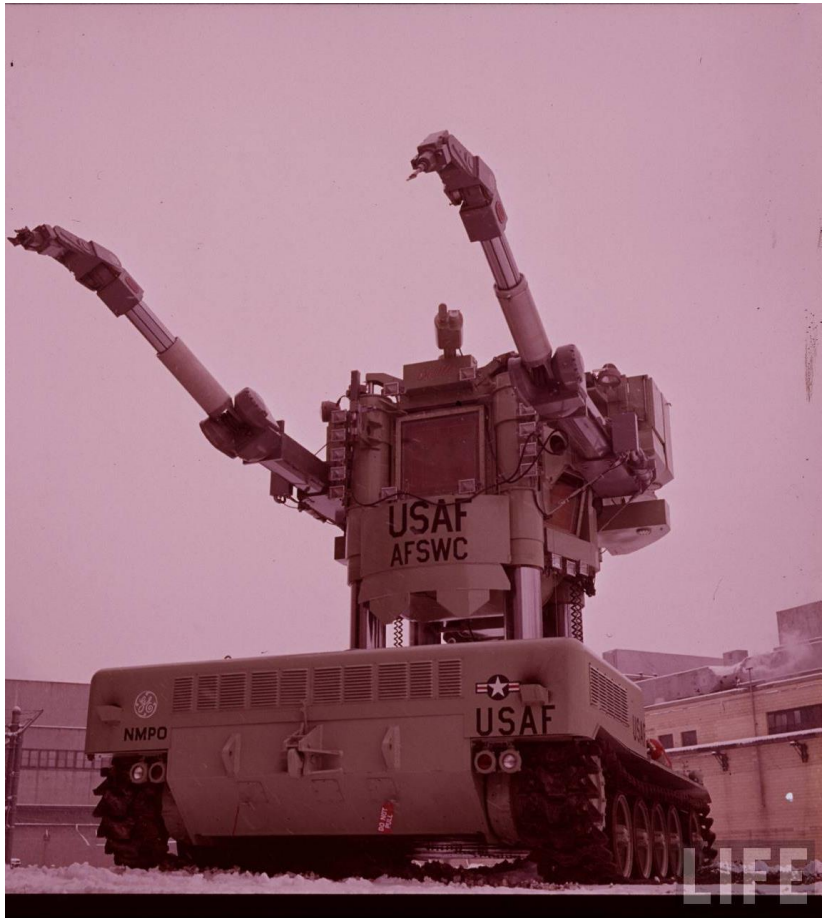
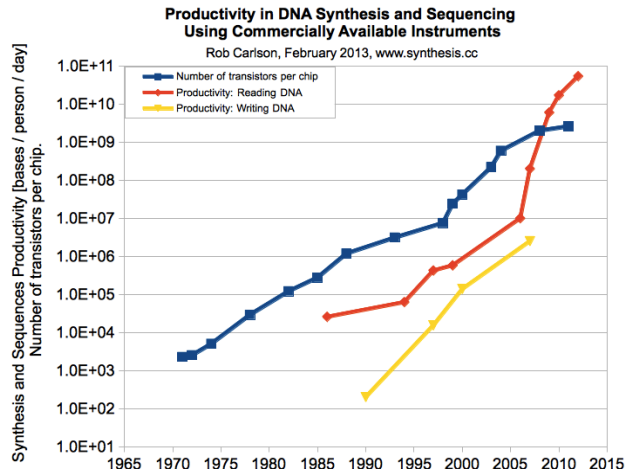


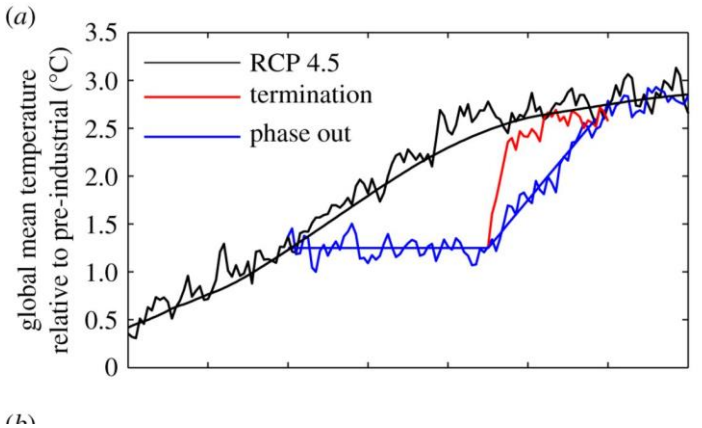
Figure 4. Surface air temperature changes for the 150 Tg case averaged for June, July, and August of the year of smoke injection and the next year. Effects are largest over land, but there is substantial cooling over oceans, too. The warming over Antarctica in year 0 is for a small area, is part of normal winter interannual variability, and is not significant. Also shown as red circles are two locations in Iowa and Ukraine, for which time series of temperature and precipitation are shown in Figures 5 and 7.

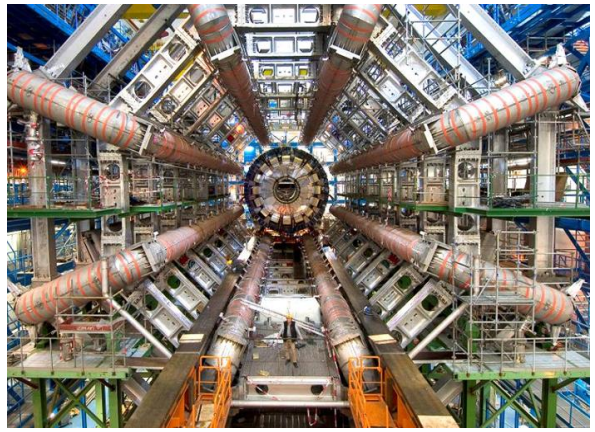
Alan Robock, Luke Oman, and Georgiy L. Stenchikov, Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, D13107, doi:10.1029/2006JD008235, 2007

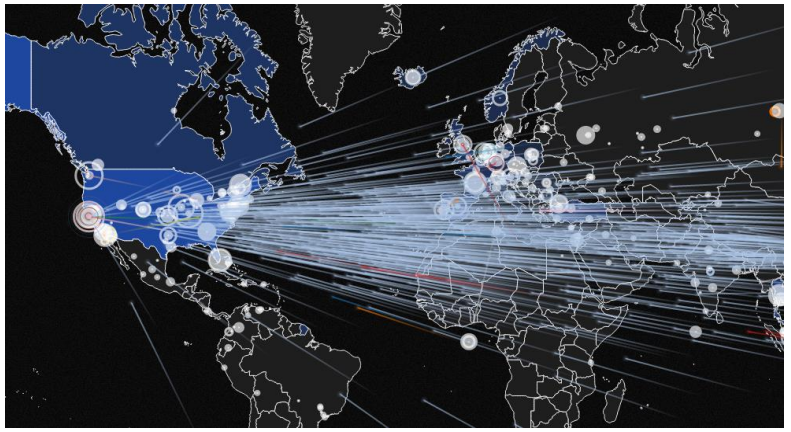
Technological risks

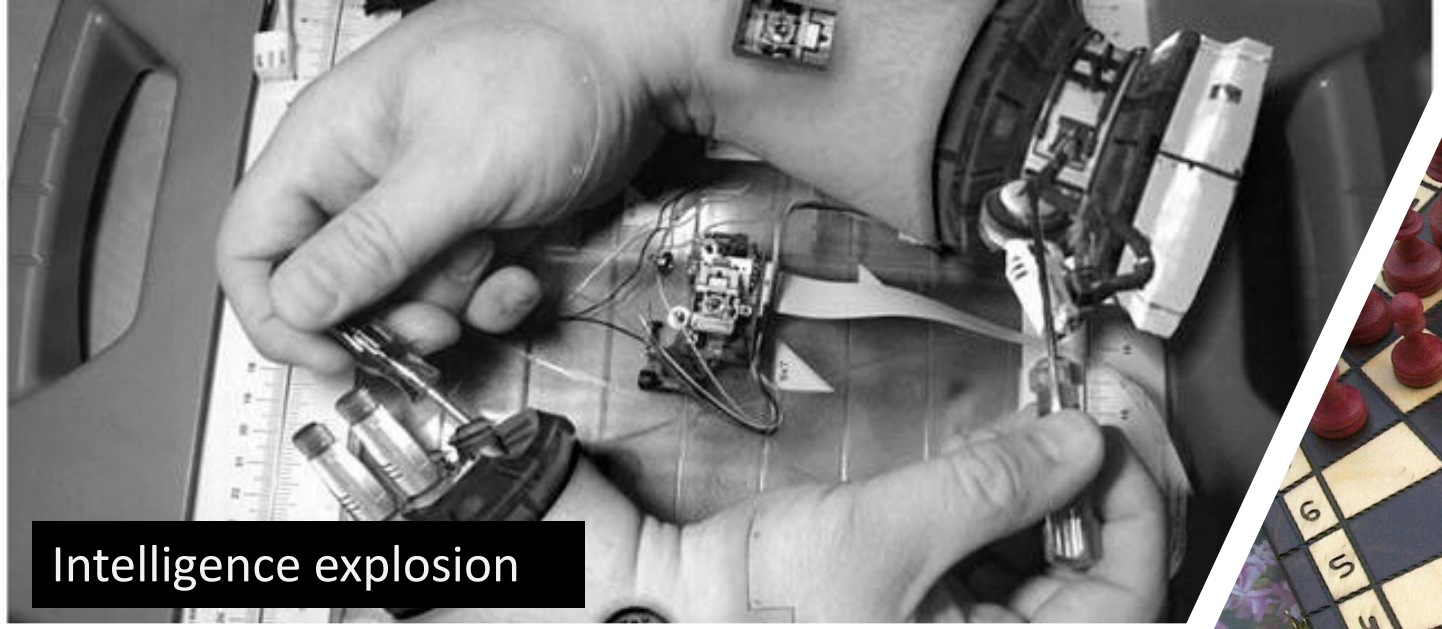








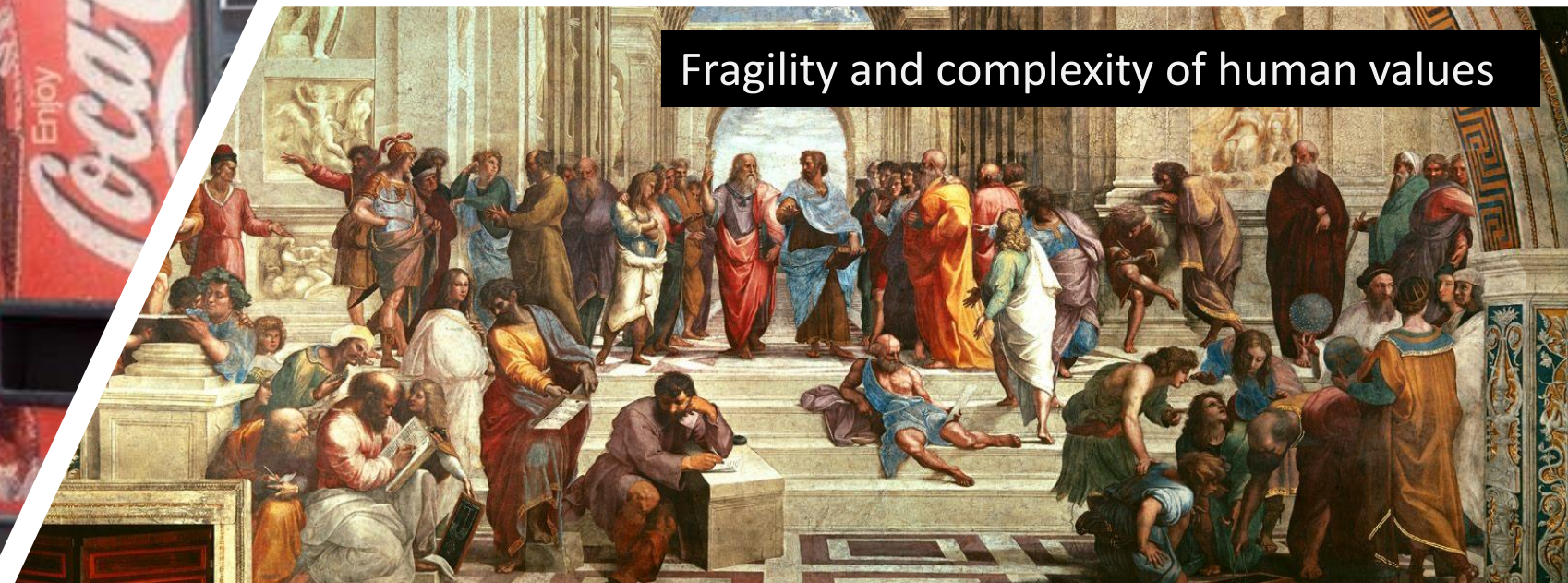




Intelligence explosion

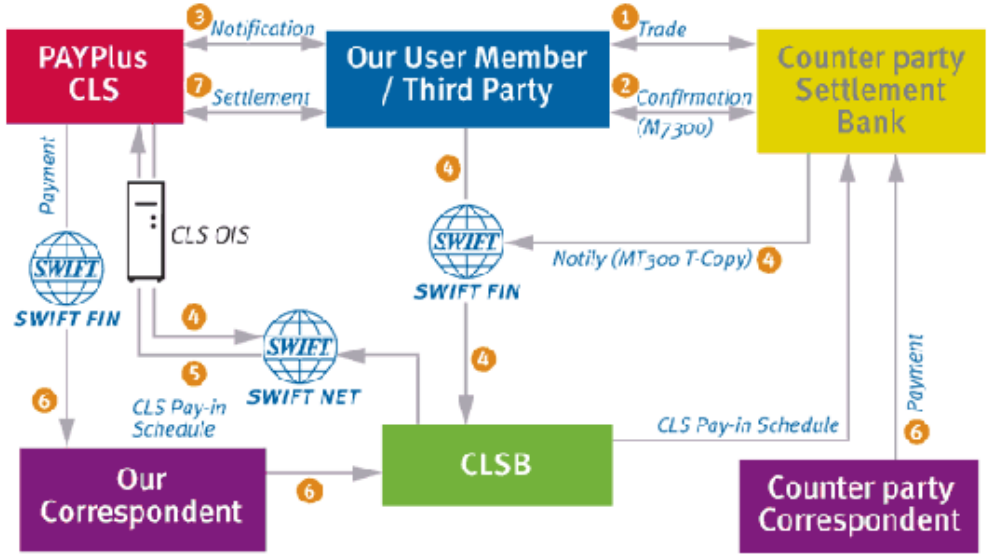
Orthogonality between intelligence and goals

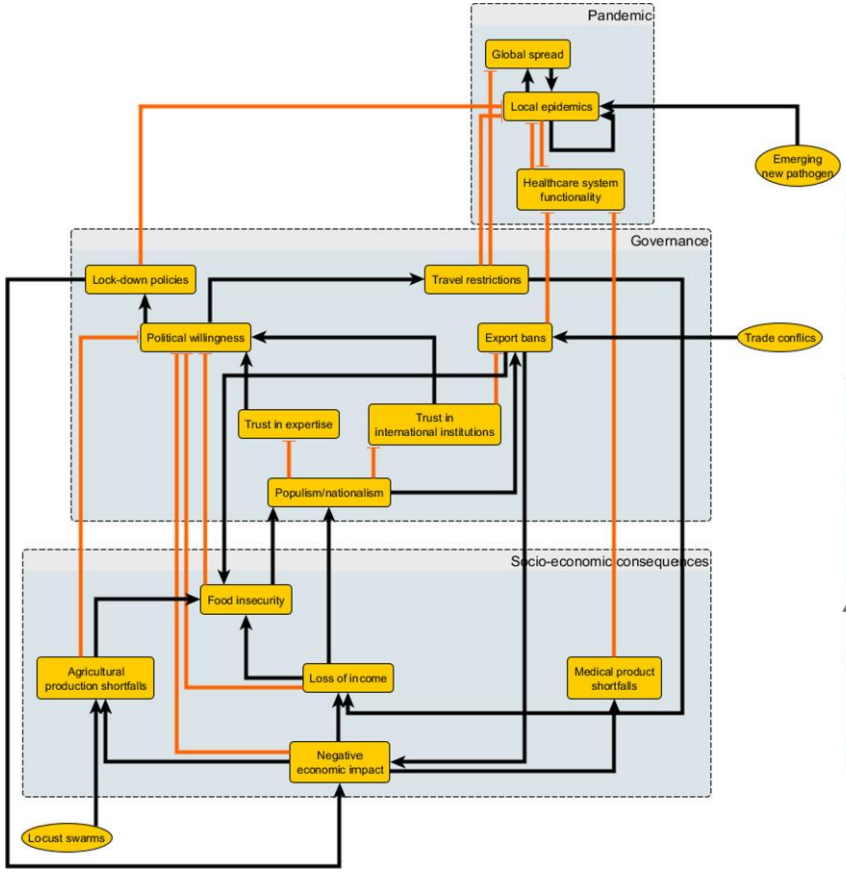
Convergent instrumental goals

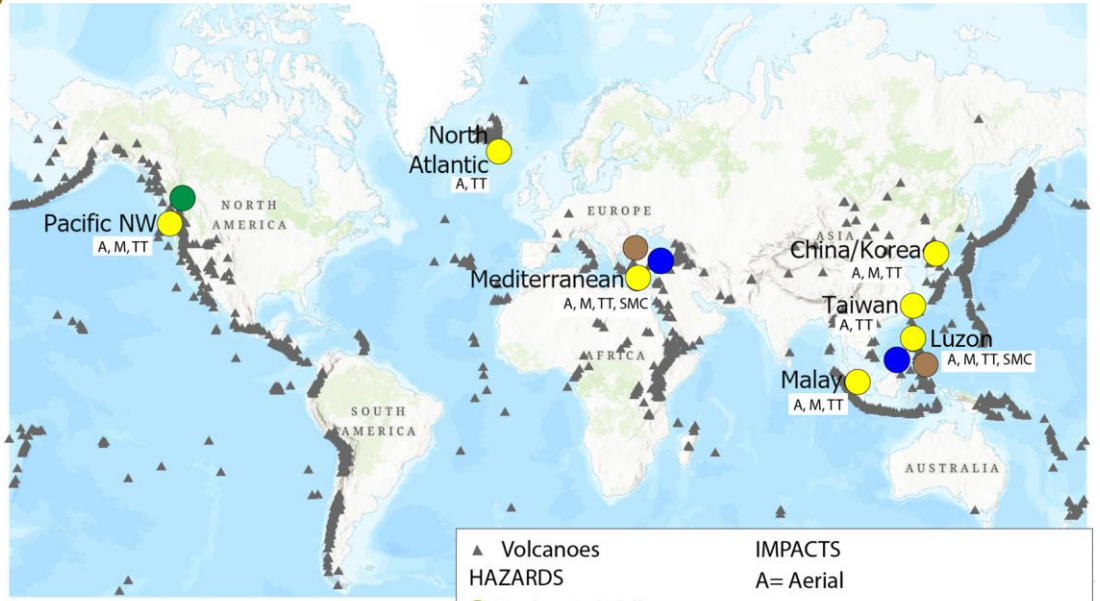


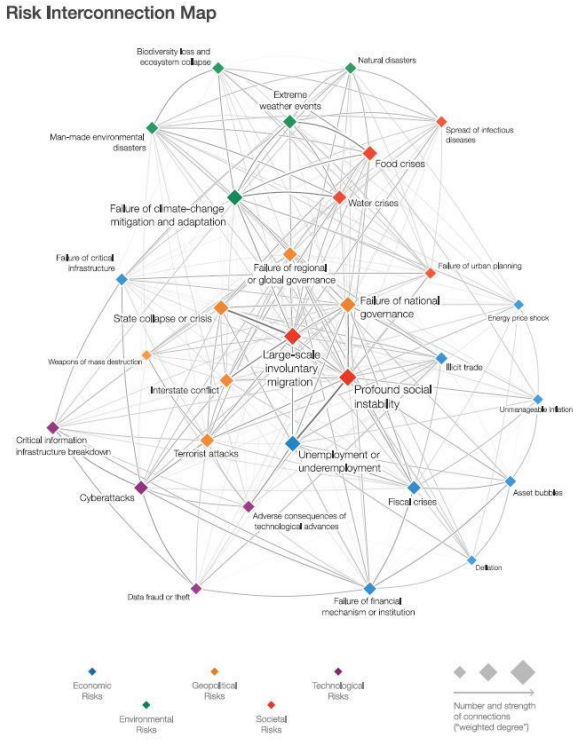
Fragility and complexity of human values

Systemic risks









How likely?

Toby Ord:

Existential catastrophe via	*Chance within the next 100 years*
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	**~ 1 in 10,000**
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
'Naturally' arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	**~ 1 in 6**
Total existential risk	**~ 1 in 6**

When and how did we learn about them?

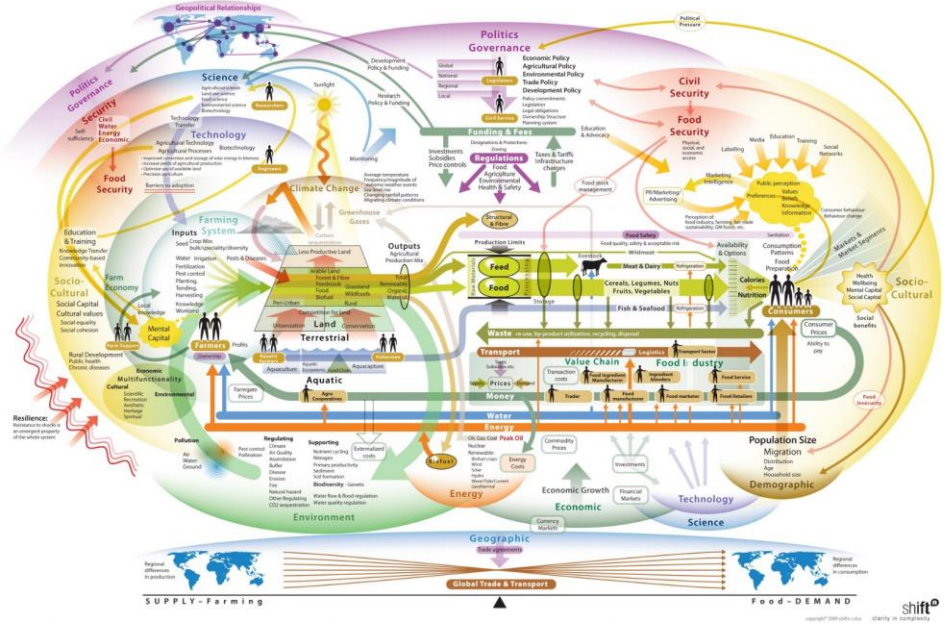
- We are finding more and more
  - Ancient: war, famine, plague, volcanism, earthquakes/tsunamis
  - Scientific age: impacts, supernovas, species extinction, ice ages
  - Early 20th century: nuclear weapons, environmental destruction, civilizational traps
  - Recent decades: Supervolcanism, geomagnetic storms, climate, AI, biotech, nanotech, systemic, physics risks, climate engineering failure…
- Not just risk aversion
- Even recognizing that they are a thing took surprisingly long!
- We are not actively searching for them (yet)

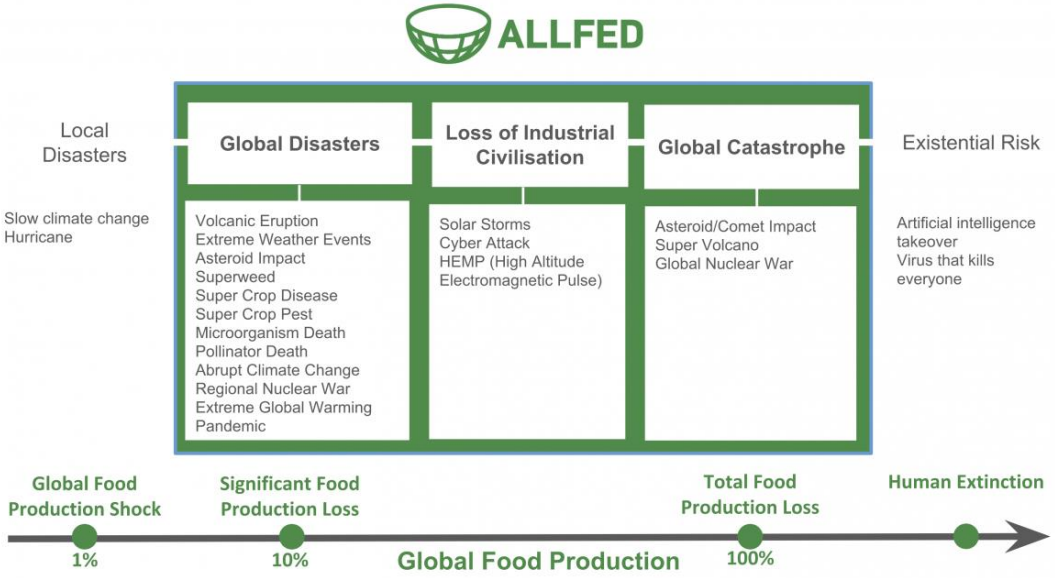
What do we know?

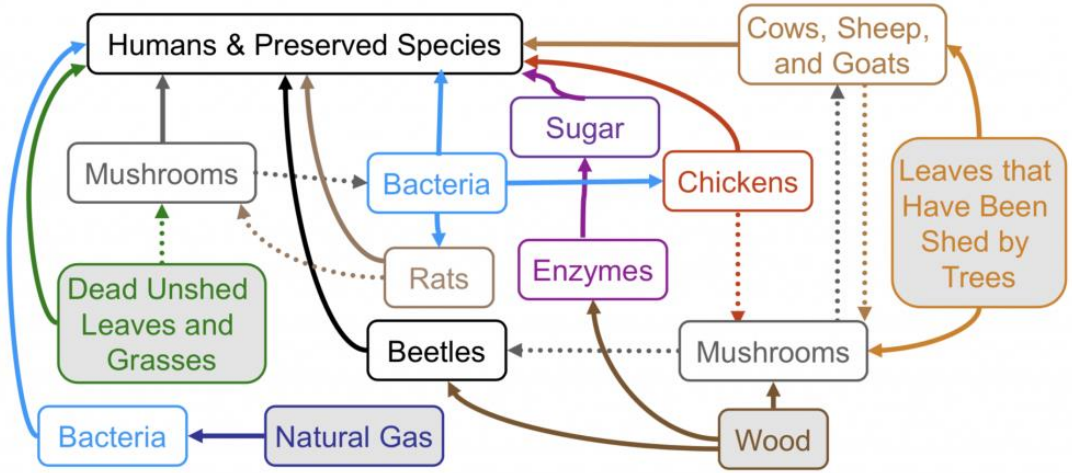
- Natural GCR/xrisks: small total risk, hard to avoid; natural pandemics dominate
- Anthropogenic risks dominate: higher likelihood and severity
- Technological risks are far more unpredictable (and potentially avoidable with design-ahead)
- Systemic risks combine other risks
- With proper mitigation and resiliency many GCRs are survivable

The food cascade

- Much of the total harm from nuclear wars, supervolcanic eruptions, meteor impacts, some biological risks and systemic risks comes from global agricultural collapse.
- Solution: rapid shift to alternative foods







What to do?

- Reduce risk that bad things happen
  - Pursue technology more carefully: manage information hazards, oversight, mandate safety features, change order of technology arrival
  - Gather information, set priorities
  - Set up prevention: vaccination, treaties, failsafes, computer security…
- Reduce their impact
  - Early warning and response: monitoring systems, faster responses, better coordination, better containment, …
  - Prevention systems: health care, cybersecurity
  - Prevent risks to defence mechanisms
  - Artificial immune systems
  - Alternative food
- Make sure we can recover if something goes wrong
  - More resources, trust, diversity
  - Spread: Refuges (natural and deliberate), space