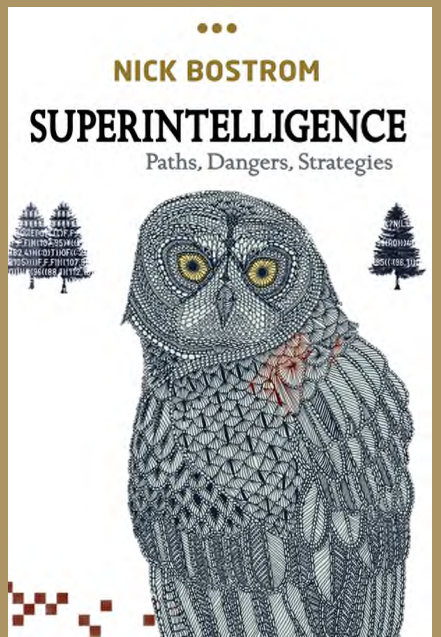
Crucial Considerations and Wise Philanthropy

Professor Nick Bostrom
Director, Future of Humanity Institute
Oxford Martin School
Oxford University



What is a crucial consideration?

a consideration such that if it were taken into account it would overturn the conclusions we would otherwise reach about how we should direct our efforts

an idea or argument that might plausibly reveal the need for not just some minor course adjustment in our practical endeavors but a major change of direction or priority

A crucial consideration (CC) is a consideration that radically changes the expected value of pursuing some high-level subgoal.

Related concepts

- Crucial consideration components
  - Considerations (arguments, ideas, data) which, while not on their own amounting to CCs, have a substantial probability of serving a central role within a CC
- Deliberation ladders
  - Sequence of CCs

Should I vote in the national election?

(A1) I should vote in order to put the better candidate into office.

(A2) My vote is extremely unlikely to make a difference. I should not vote but put my time to better use.

(A3) Although it is unlikely that my vote would make a difference, the stakes are very high: millions of lives are affected by the president. So even if the chance of my vote being decisive is only one in several millions, the expected benefit is large enough to be worth a trip to the polling station.

(A4) If the election is not close, my vote will make no difference. If the election *is* close, approximately half the votes will be for the wrong candidate—implying that either the candidates are of almost exactly equal merit (and it scarcely matters who wins) or a typical voter's judgment of the candidates' merit is *extremely* unreliable and carries almost no signal. I should not bother to vote.

(A5) I am a much better judge of the candidates' merits than is the typical voter. I should vote.

(A6) Psychological studies show that people tend to be overconfident: almost everybody believes themselves to be above average, but they are as likely to be wrong as right about that. If I'm as likely to vote for the wrong candidate as is the typical voter, then my vote would add negligible information to the selection process. I should not vote.

Should I vote in the national election? (cont…)

(A7) The fact that I have gone through the previous six steps demonstrates that I am exceptionally savvy. I'm therefore more likely to pick the best candidate. I should vote.

(A8) If I'm really so special, then the opportunity cost of my going to the polling booth is especially high. I should not vote but instead devote my rare abilities to some higher-value activity.

(A9) If I don't vote, my acquaintances will see that I have failed to support the candidate both they and I think is best. This could make me look weird or disloyal, diminishing my influence (which I would otherwise have used for good ends). I should vote.

(A10) It is important to stand up for one's convictions. It stimulates fruitful discussion. Moreover, when I explain the intricate reasoning that led me to refrain from voting, my friends might think I'm clever. I should not vote.

(A11) …

Should we favour more funding for X-Tech research?

(B1) We should fund X-Tech research because there are many potential future applications in medicine, manufacturing, clean energy, etc.

(B2) But X-Tech would also have important military applications. It might ultimately make it possible to produce new kinds of weapons of mass destruction that would pose a major existential risk. We should *not* fund it.

(B3) If this kind of X-Tech is possible, it will almost certainly be developed sooner or later even if *we* decide not to pursue it. If responsible people refrain from developing it, it would be developed by irresponsible people, which would make the risks even greater. We should fund it.

(B4) But *we* are already ahead in its development. Extra funding would only get us there sooner—leaving us with less time to properly prepare for the dangers. So we should not add funding.

(B5) Look around: you'll see virtually no serious effort to prepare for the dangers of X-Tech. This is because serious preparation will begin only *after* a massive project is already underway to develop X-Tech—only then will people will take the prospect seriously. The earlier such a project is initiated, the longer it will take to complete (since it will be starting from a lower general level of technological capability). Launching the serious project now therefore means more time for serious preparation. So we should push on as hard as we can.

Should we favour more funding for X-Tech research? (cont…)

(B6) The level of risk will be affected by other factors than the amount of serious preparation that has been made specifically to counter the threat from X-Tech. For instance, machine superintelligence or ubiquitous surveillance might be developed before X-Tech, eliminating or mitigating the risks of the latter. Although these other technologies may pose grave risks of their own, those risks would have to be faced *anyway*, and X-Tech would not reduce them. So the preferred sequence is that we get superintelligence or ubiquitous surveillance before we get X-Tech. So we should oppose extra funding for X-Tech.

(B7) However, if we oppose extra funding for X-Tech, the people working in X-Tech will dislike us; and other scientists might regard us as being anti-science. This will reduce our ability to work with these scientists, hampering our efforts on more specific issues, efforts that stand a much better chance of making a material difference than any attempts on our part to influence the level of national funding for X-Tech. So we should not oppose extra funding for X-Tech.

(B8) …

Why is utilitarianism rich in CCs?

- Knowledge & shaping
  - We have more knowledge and experience of human life at the personal level.
- Difficulties in understanding the goal itself
  - We find it difficult to grasp the kinds of utility functions imputed by utilitarianism (and some versions of egoism)
- Semi-nearness to historical pivot point
  - If we stand in the vicinity to a pivot point of history, we may have special opportunities to influence the long-term future.
- Recent discovery of key exploration tools
  - We have recently discovered some key concepts and ideas that may unlock further important discoveries

Evaluation function

$Eval_{chess}$ =

(c_1×material) + (c_2×mobility) + (c_3×king safety) + (c_4×center control) + ...

Evaluation function

$Eval_{chess}$ =

$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

Evaluation function

$Eval_{chess}$ =

$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

$Eval_{moral\_goodness}$ = ?

Evaluation function

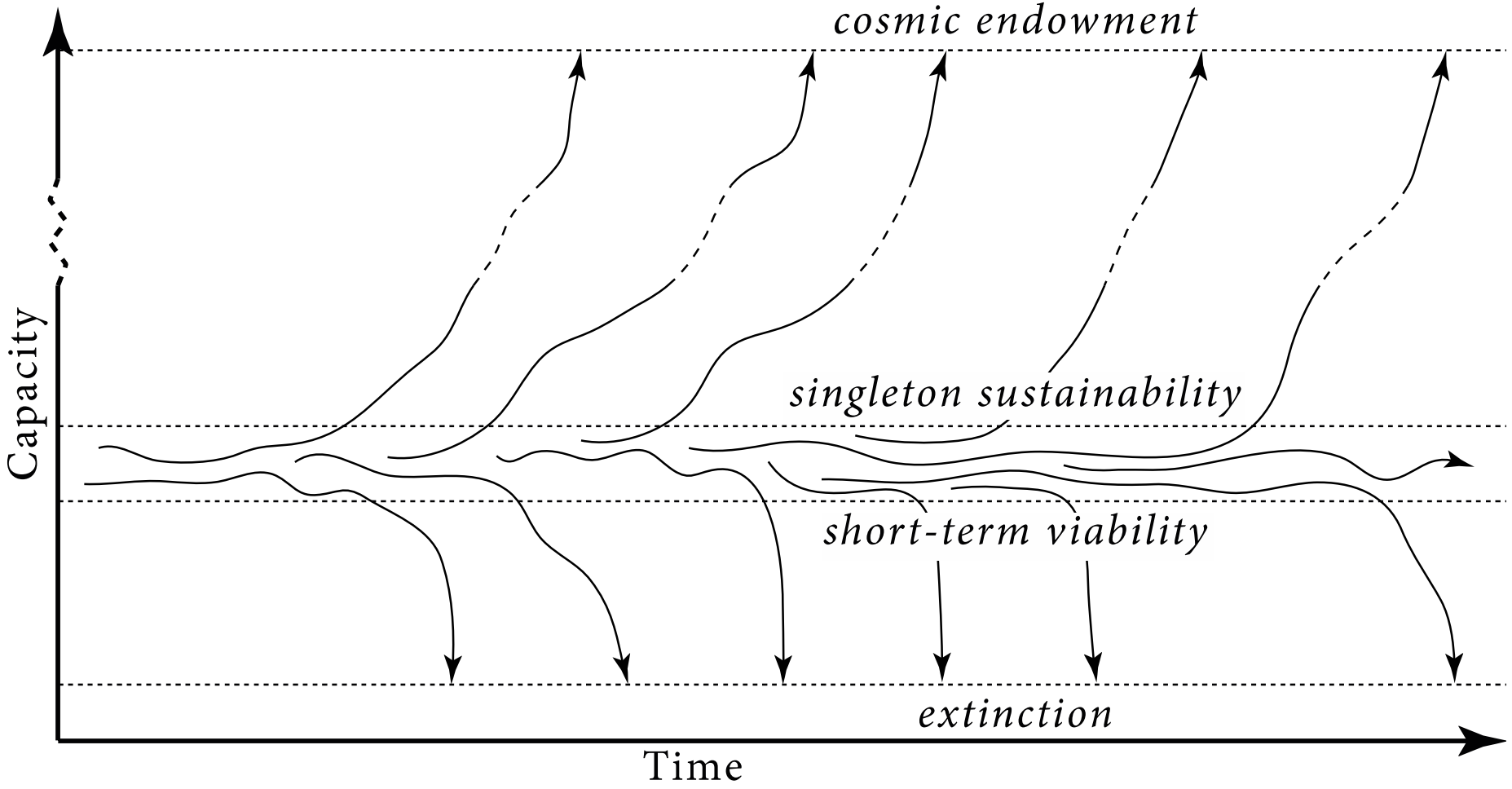
$Eval_{chess}$ =

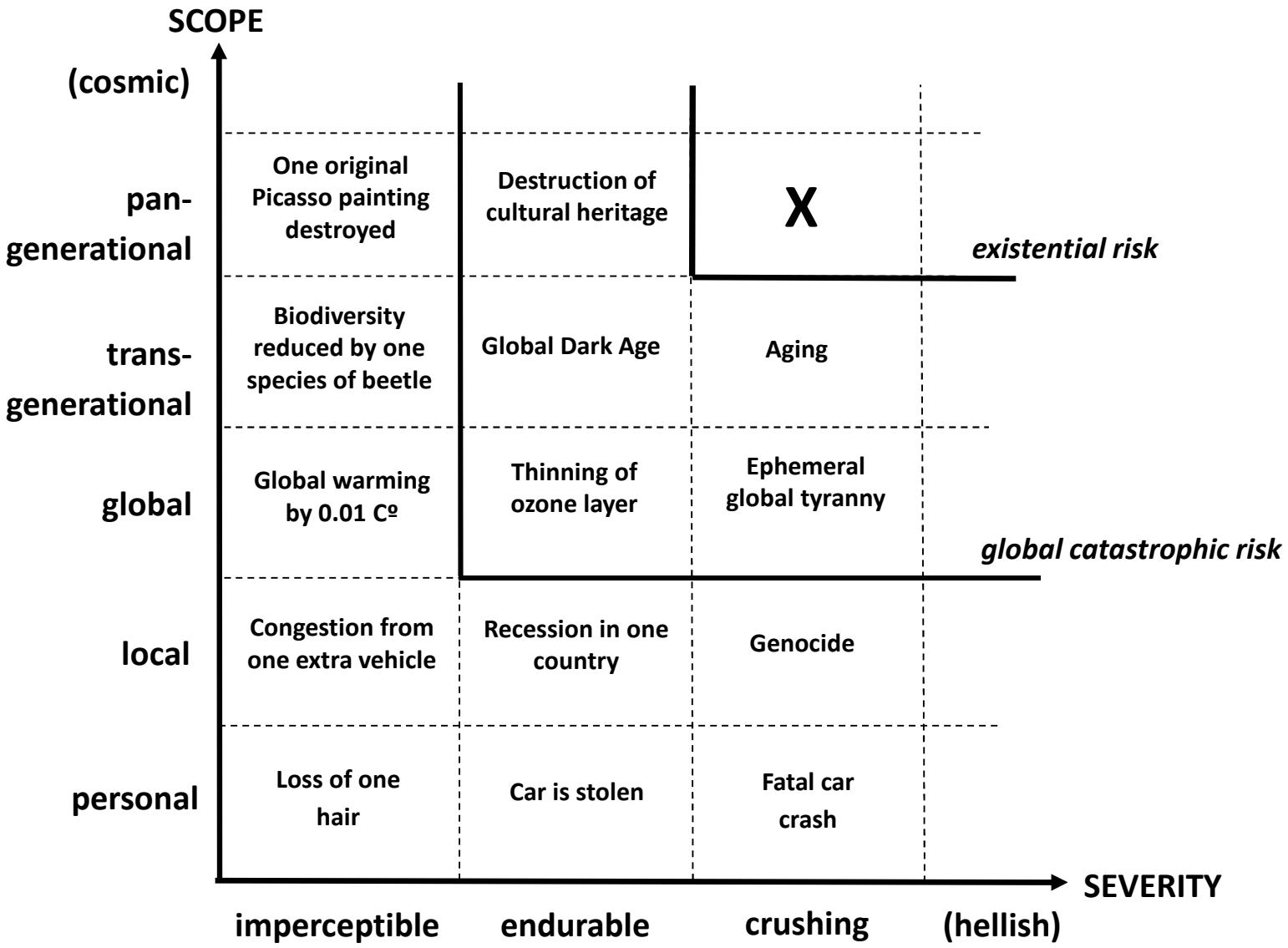
$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

$Eval_{utilitarian}$ = ?


cosmic endowment
singleton sustainability
short-term viability
extinction
Capacity
Time


SCOPE
(cosmic)
pan-generational
trans-generational
global
local
personal
One original Picasso painting destroyed
Destruction of cultural heritage
X
existential risk
Biodiversity reduced by one species of beetle
Global Dark Age
Aging
Global warming by 0.01 Cº
Thinning of ozone layer
Ephemeral global tyranny
global catastrophic risk
Congestion from one extra vehicle
Recession in one country
Genocide
Loss of one hair
Car is stolen
Fatal car crash
SEVERITY
imperceptible
endurable
crushing
(hellish)

MAXIPOK

Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe

argmax [- P(existential catastrophe / action)]

Evaluation function

$Eval_{chess}$ =

$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

$Eval_{utilitarian} \approx$ MAXIPOK

Evaluation function

$Eval_{chess}$ =

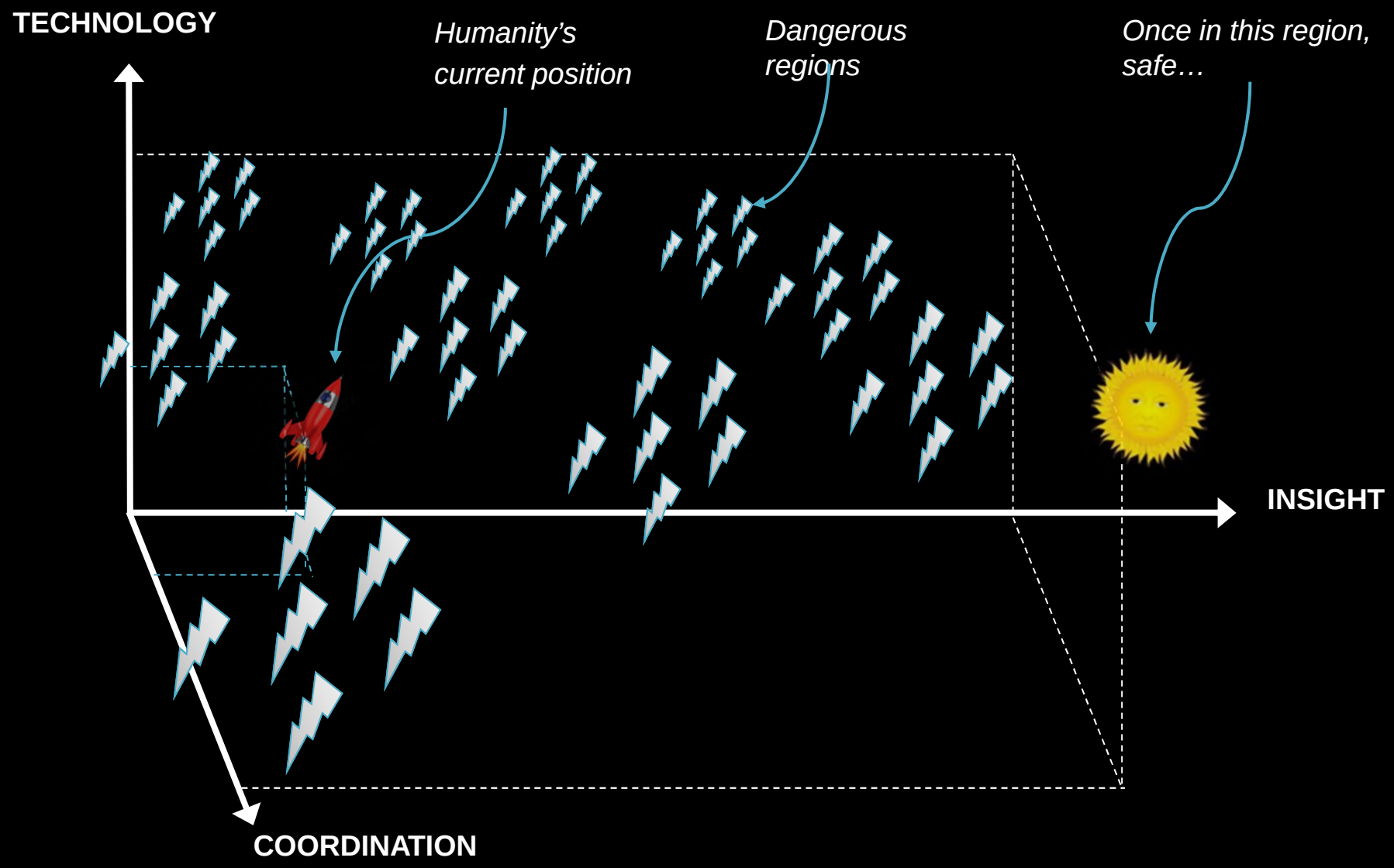
$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

$Eval_{maxipok}$ = ?

Dynamic sustainability



Evaluation function

$Eval_{chess}$ =

$(c_1 \times material) + (c_2 \times mobility) + (c_3 \times king\ safety) + (c_4 \times center\ control) + ...$

$Eval_{public\_policy}$ =

$(c_1 \times GDP) + (c_2 \times employment) + (c_3 \times equality) + (c_4 \times environment) + ...$

$Eval_{maxipok}$ = f (wisdom, coordination, differential tech development, …)

Principle of differential technological development

Retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies.

Cause selection vs. signature determination

- Causes should be high-leverage
- Signposts should be useful for general orienteering
  - we should have strong reason to think we know their directionality
  - they should ideally be visible from afar

Some (very) tentative signposts

- Computer hardware?——————————————No
- Whole brain emulation?——————————————No(?)
- Biological cognitive enhancement?——————Yes
- Artificial intelligence?——————————————No
- Lead of AI frontrunner?-——————————————---Yes
- Solutions to the control problem?——————Yes
- Effective altruism movement?——————————-Yes
- International peace and cooperation?————Yes
- Synthetic biology?-——————————————-No(?)
- Nanotechnology?-——————————————--No
- Economic growth?-——————————————-- ?
- Small and medium-scale catastrophe prevention?-- ?

List of some areas with candidate remaining CCs or CCCs

- Counterfactual trade
- Simulation stuff
- Infinite paralysis
- Pascalian muggings
- Different kinds of aggregative ethics (total, average, negative)
- Information hazards

- Aliens
- Baby universes
- Other kinds of moral uncertainty
- Other game theory stuff

- Pessimistic metainduction; epistemic humility; anthropics
- Insects, subroutines

Some partial remedies

- don’t act precipitously (and in ways that are irrevocable)
- invest more in analysis (find and assemble CCs)
- take into account that EV-changes are probably smaller than they appear (quiescence search, meta stuff)
- use parliamentary / mixture models
- focus more on near term & convenient objectives (e.g. if one is partly egoist and partly altruist, but on altruism one is on a deliberation ladder, then maybe go with the egoistic part)
- invest in developing inner capacity: not more powers but rather propensity to use powers better

Some (very) tentative signposts

- Computer hardware?——————————————No
- Whole brain emulation?——————————————No(?)
- Biological cognitive enhancement?——————Yes
- Artificial intelligence?——————————————No
- Lead of AI frontrunner?-——————————————---Yes
- Solutions to the control problem?——————Yes
- Effective altruism movement?——————————-Yes
- International peace and cooperation?————Yes
- Synthetic biology?-——————————————-No(?)
- Nanotechnology?-——————————————--No
- Economic growth?-——————————————-- ?
- Small and medium-scale catastrophe prevention?-- ?

Technological completion conjecture

If scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained.

Need for speed?

“I instinctively think go faster. Not because I think this is better for the world. Why should I care about the world when I am dead and gone? I want it to go fast, damn it! This increases the chance I have of experiencing a more technologically advanced future.”

— the blog-commenter “washbash”

The risk of creativity

The risk of creativity

Hazardous future techs?

- Machine intelligence
- Synthetic biology
- Molecular nanotechnology
- Totalitarianism-enabling technologies
- Human modification
- Geoengineering
- Unknown
- Unknown
- Unknown
- Unknown

Past philanthropy

- Share meat with the tribe?
- Hold a festival for the people?
- 347 B.C.Plato's will left his farm to a nephew with instructions the proceeds be used to support students and faculty at the academy he founded.

notes

- CCCs
- Def deliberation ladder