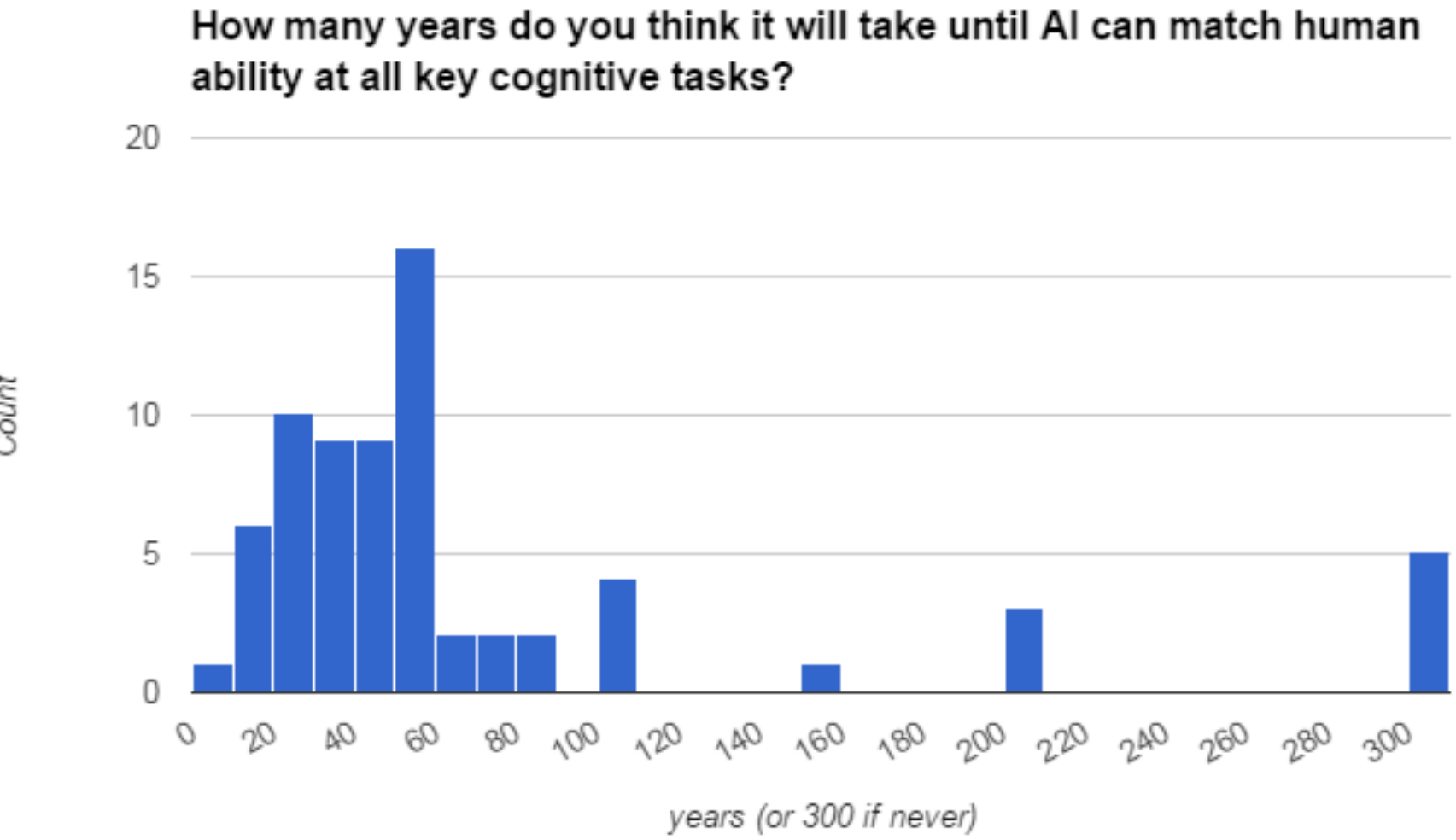
The Road Ahead

Nick Bostrom

Professor, Oxford University
Director, Future of Humanity Institute
Director, Programme on the Impacts of Future

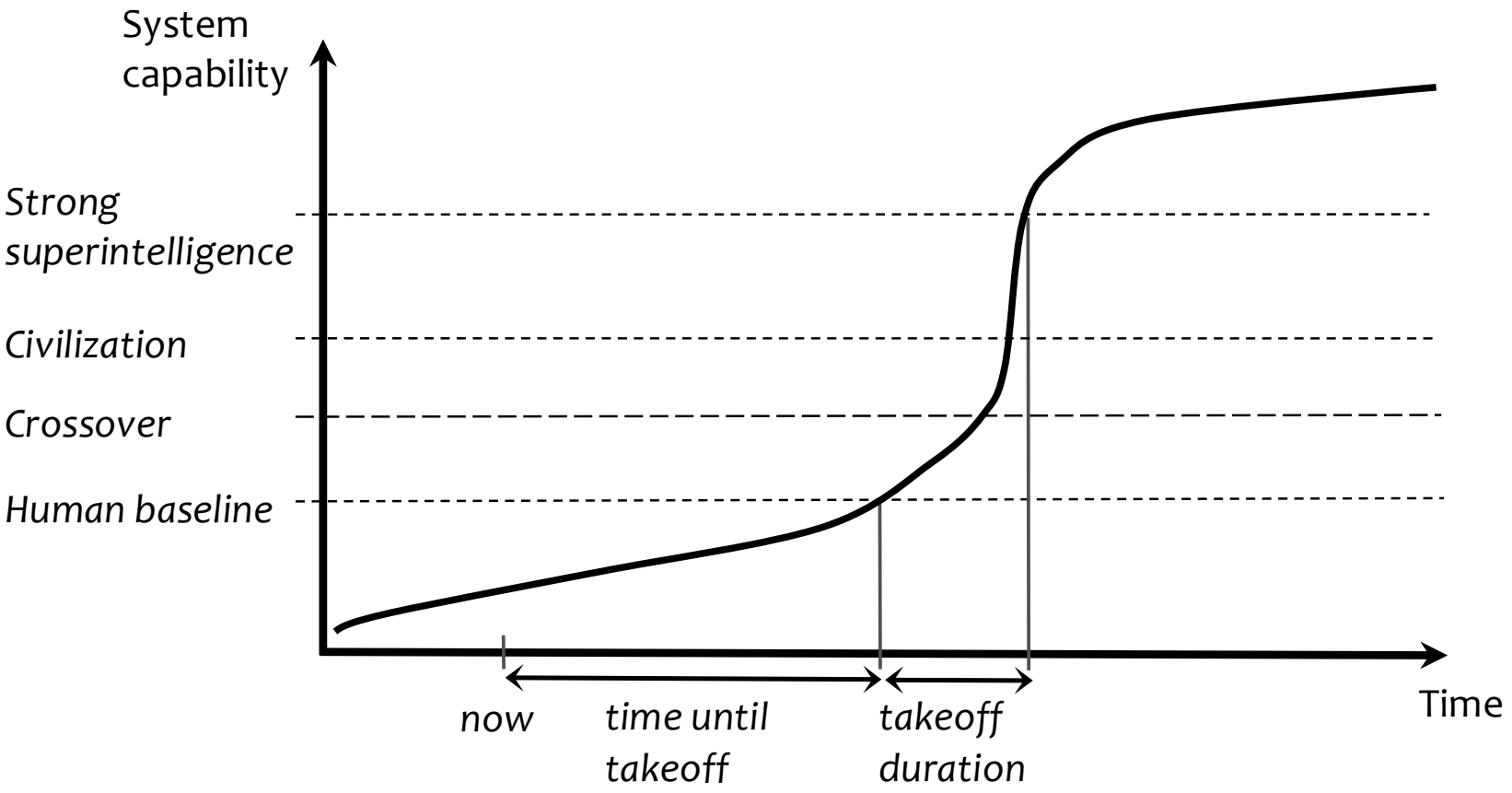
Future of Humanity Institute
UNIVERSITY OF OXFORD

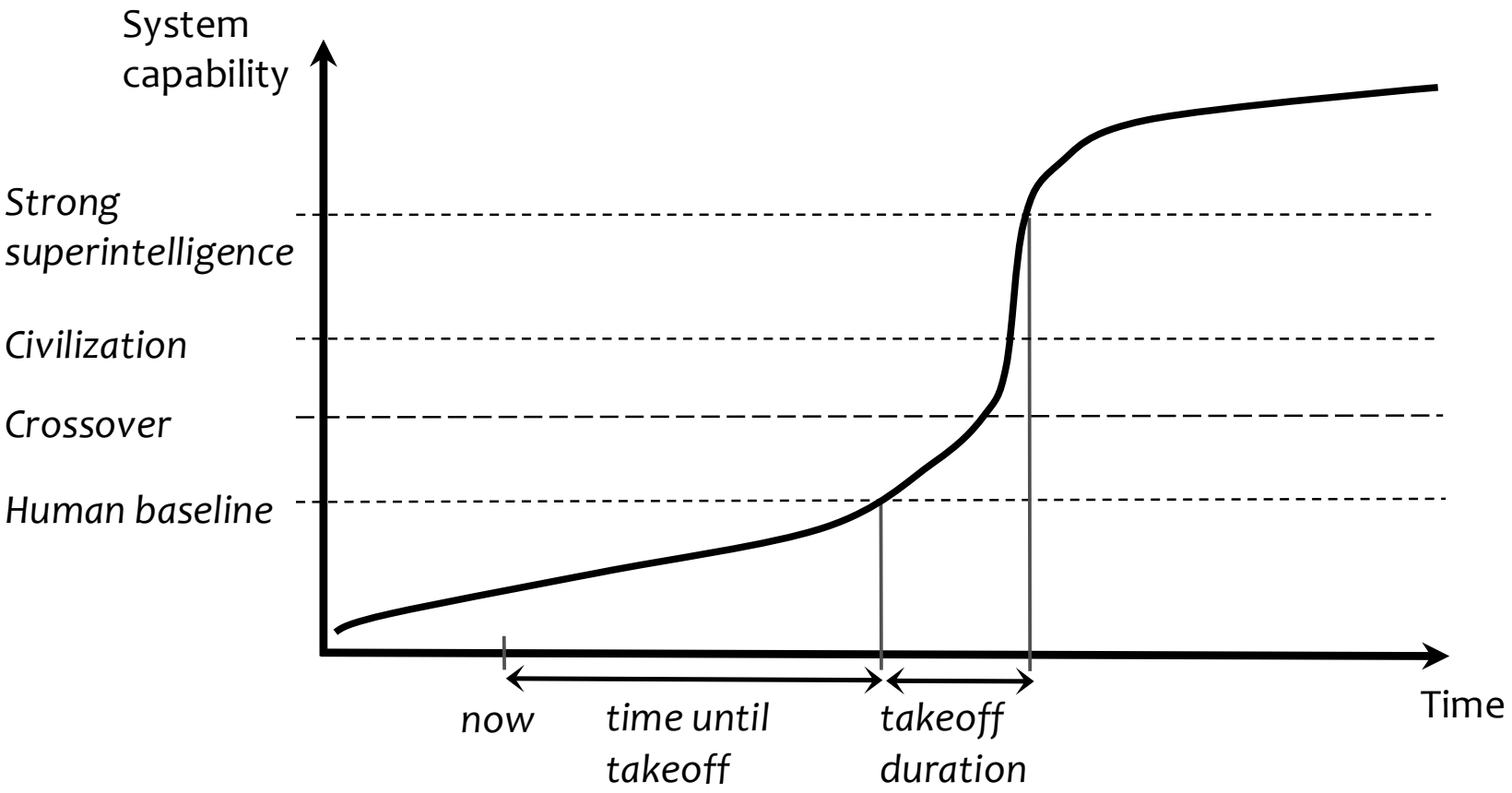
This conference



When will HLMI be achieved?

	10%	*50%*	*90%*
PT-AI	***2023***	***2048***	***2080***
AGI	***2022***	***2040***	***2065***
EETN	***2020***	***2050***	***2093***
TOP100	***2024***	***2050***	***2070***
Combined	***2022***	***2040***	***2075***

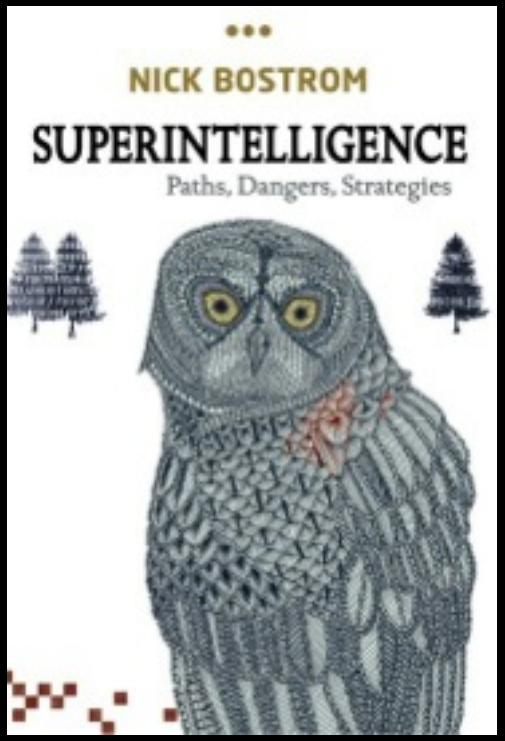

System capability
Strong superintelligence
Civilization
Crossover
Human baseline
now
time until takeoff
takeoff duration
Time



Fast – minutes, hours, days
Slow – decades, centuries
Intermediate – months, years

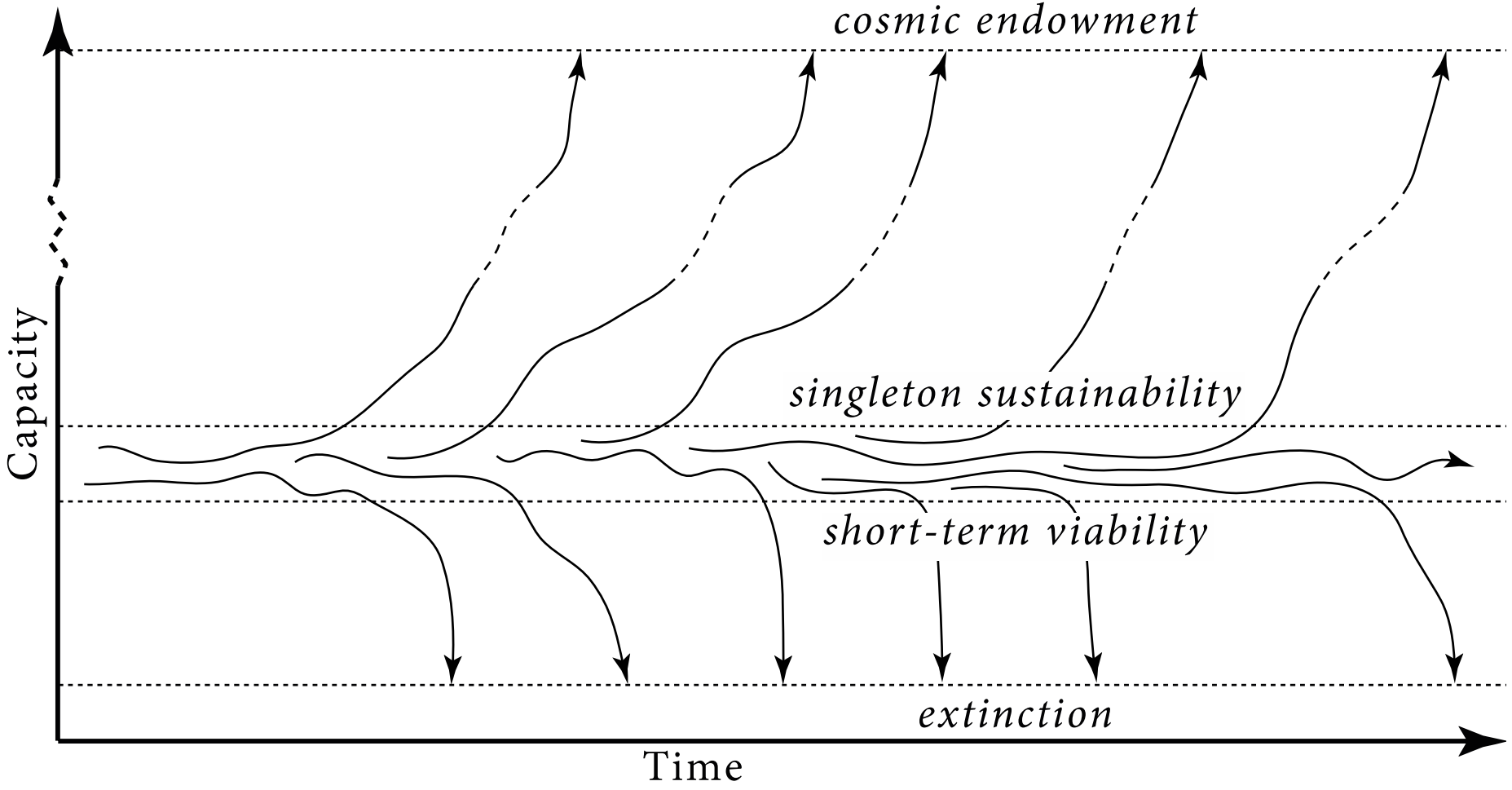
In the book

- Transition to machine intelligence era pivotal event
- Superintelligence could be extremely powerful
- Takeoff *might* be rather fast
- Singleton or multipolar outcome
- In either case, things could easily go wrong
- In non-obvious ways



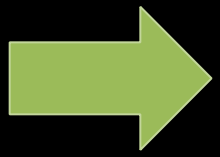
- Wire-heading
- Mind crime
- Emergent agentic behavior in powerful optimization processes
- Convergent instrumental reasons
- Simulations, blackmail
- …

- Step 1: Recognize that a problem exists
- Step 2: Recognize that the problem is non-trivial


cosmic endowment
singleton sustainability
short-term viability
extinction
Capacity
Time

- Step 1: Recognize that a problem exists
- Step 2: Recognize that the problem is non-trivial
- Step 3: Care sufficiently about what happens

- Step 1: Recognize that a problem exists
- Step 2: Recognize that the problem is non-trivial
- Step 3: Care sufficiently about what happens to humanity's cosmic endowment



Proceed to develop and implement some wise course of action…

What needs to be done

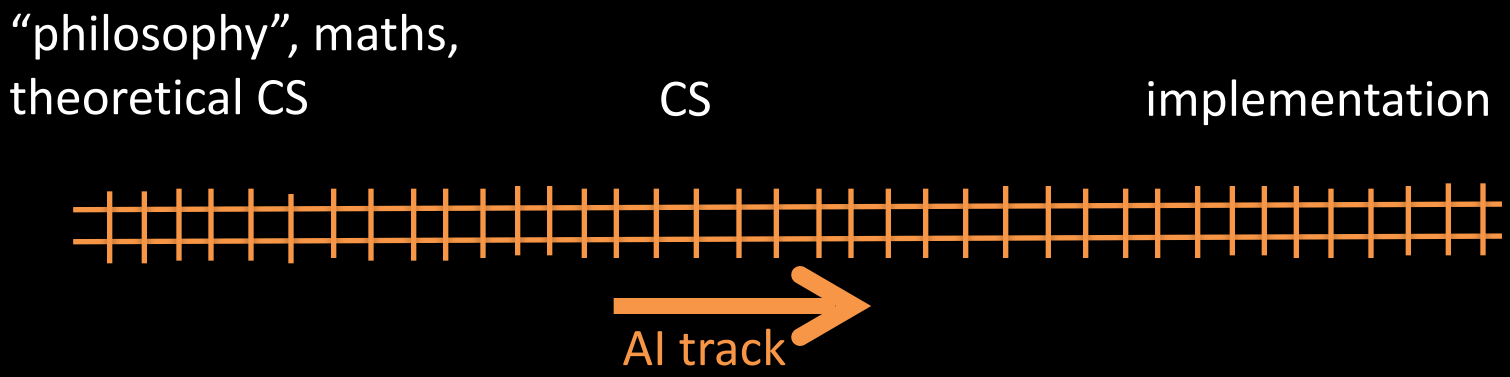
- Make theoretical progress happen on the control problem
- Build collaboration between AI developers and AI safety community to strengthen mutual understanding and trust
- Create hirable safety experts
- Ensure sufficient risk awareness so that serious AGI projects want to hire safety experts
- Ensure the field is sufficiently well ordered so that merit can be ascertained
- Promote sense of ethical responsibility & look for opportunities for positive-sum trades

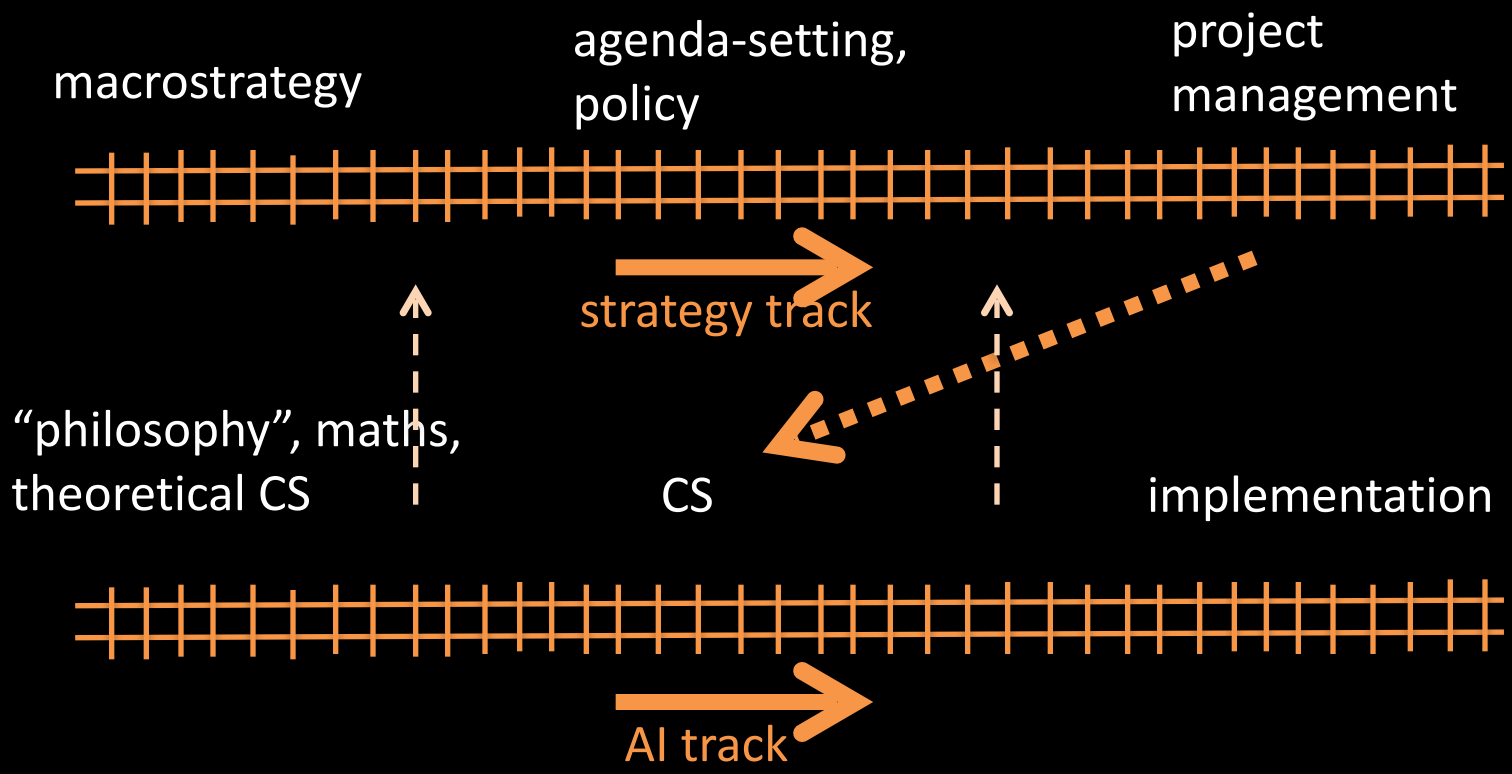
The Common Good Principle

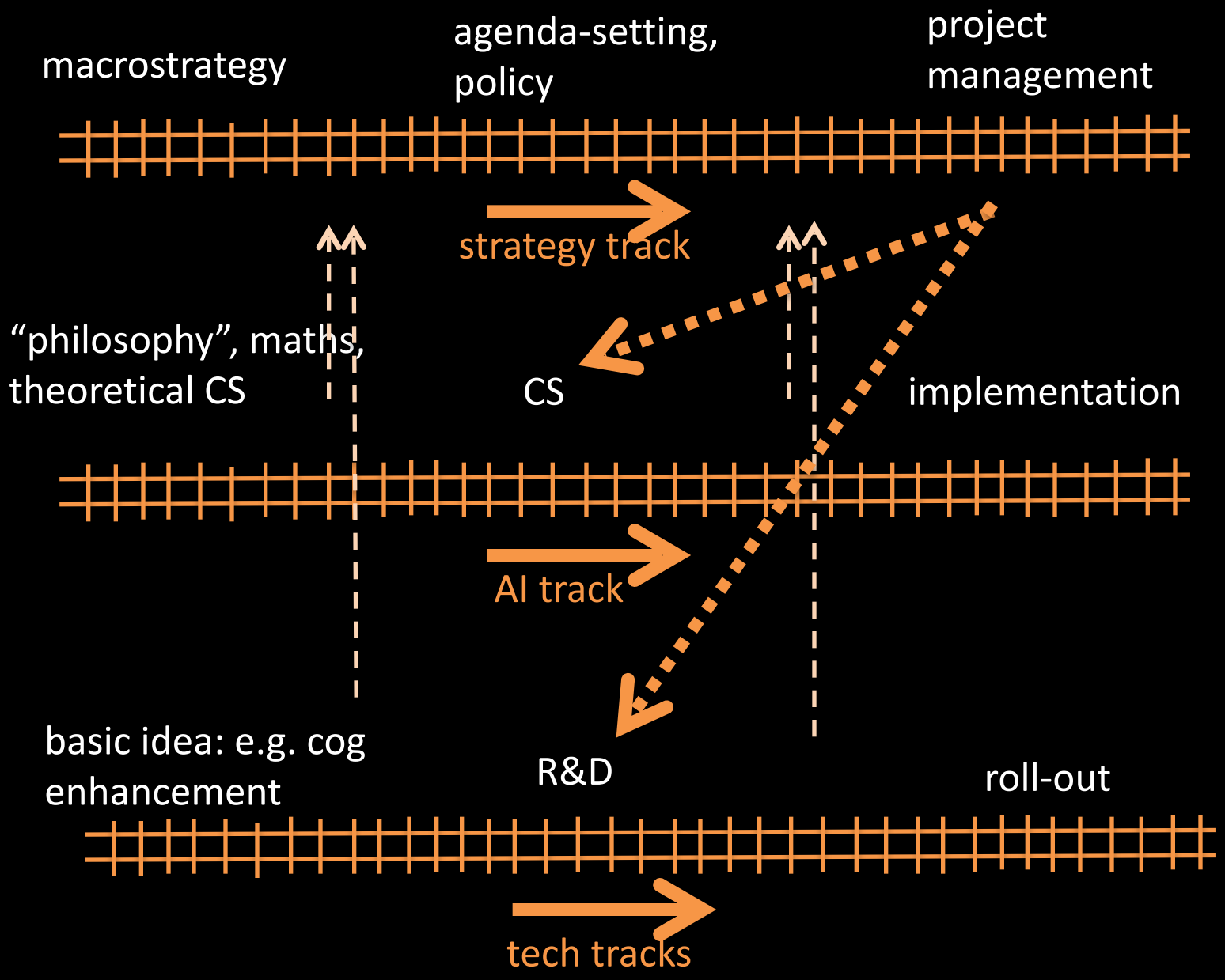
The Common Good Principle

Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals.

- is consistent with commercial rewards for work along the way


“philosophy”, maths,
theoretical CS
CS
implementation
AI track


macrostrategy
agenda-setting,
policy
project
management
strategy track
“philosophy”, maths,
theoretical CS
CS
implementation
AI track


macrostrategy
agenda-setting, policy
project management
strategy track
"philosophy", maths, theoretical CS
CS
implementation
AI track
basic idea: e.g. cog enhancement
R&D
roll-out
tech tracks

Some lines of attack

- MIRI research agenda
  - naturalized induction; ontology identification; theory of counterfactuals; logical counterfactuals; impossible possibilities; logical priors; Vingean reflection; Löbian obstacle; corrigibility; utility indifference; domesticity; multi-level world models; ambiguity identification; operator modeling; normative uncertainty
- FHI
  - tripwires (containment methods); Hail Mary; utility diversification; distillation; tool AI; oracle systems; indirect normativity; indirect approaches to specifying decision theory
- Russell
  - inverse reinforcement learning
- Christiano
  - approval-directed agents; crypt; steering problem
- Additional
  - Monitoring & diagnosis tools; concept learning; program verification; …

Research infrastructure needs (0-2 yrs)

- FHI (Oxford University)
- MIRI
- CSER (Cambridge University)
- FLI (individual researchers, events, etc.)
- e.g. Berkeley?
- e.g. MIT/Harvard?
- Industry (e.g. DeepMind, Vicarious)
- CFAR (talent scouting)


Future
of Humanity
Institute
UNIVERSITY OF OXFORD