gbs switzerland

Superintelligence

Our Final Invention

Kaspar Etter, kaspar.etter@gbs-schweiz.org
Adrian Hutter, adrian.hutter@gbs-schweiz.org

Outline

- Introduction
- Singularity
- Superintelligence
- State and Trends
- Strategy
- Summary

More Information:
superintelligence.ch

Introduction

What are we talking about?

Intelligence

«Intelligence measures an agent's ability to achieve its goals in a wide range of unknown environments.»

(adapted from Legg and Hutter)

$$\text{Intelligence} = \frac{\text{Optimization Power}}{\text{Used Resources}}$$

Universal Intelligence
arxiv.org/pdf/0712.3329.pdf

Ingredients

- **Epistemology**: Learn model of world
- **Utility Function**: Rate states of world
- **Decision Theory**: Plan optimal action

(There are still some open problems, e.g. classical decision theory breaks down when the algorithm itself becomes part of the game.)

Luke Muehlhauser: Decision Theory FAQ
lesswrong.com/lw/gu1/decision_theory_faq/

Consciousness

- … is a completely separate question!
- **Not required** for an agent to reshape the world according to its preference

Consciousness is

- **reducible** or
- **fundamental**
- and **universal**

How Do You Explain Consciousness?
David Chalmers: go.ted.com/DQJ

Machine Sentience

Open questions of immense importance:

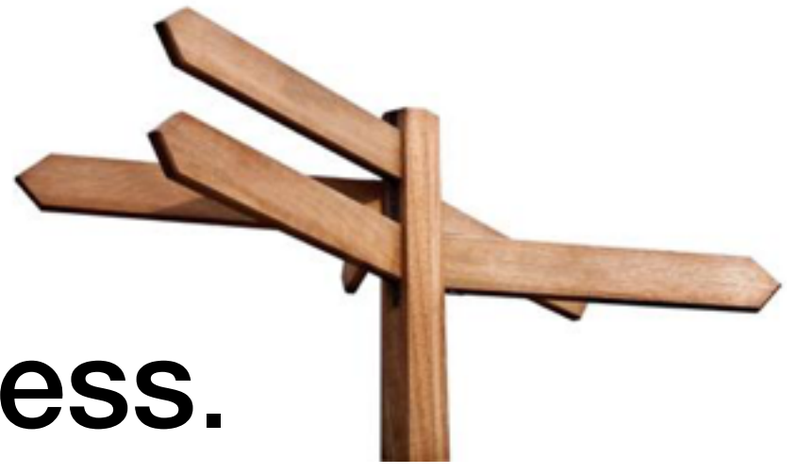
- Can simulated entities be **conscious**?
- Can machines be **moral patients**?

If yes:

- Machines deserve moral consideration
- We might live in a computer simulation

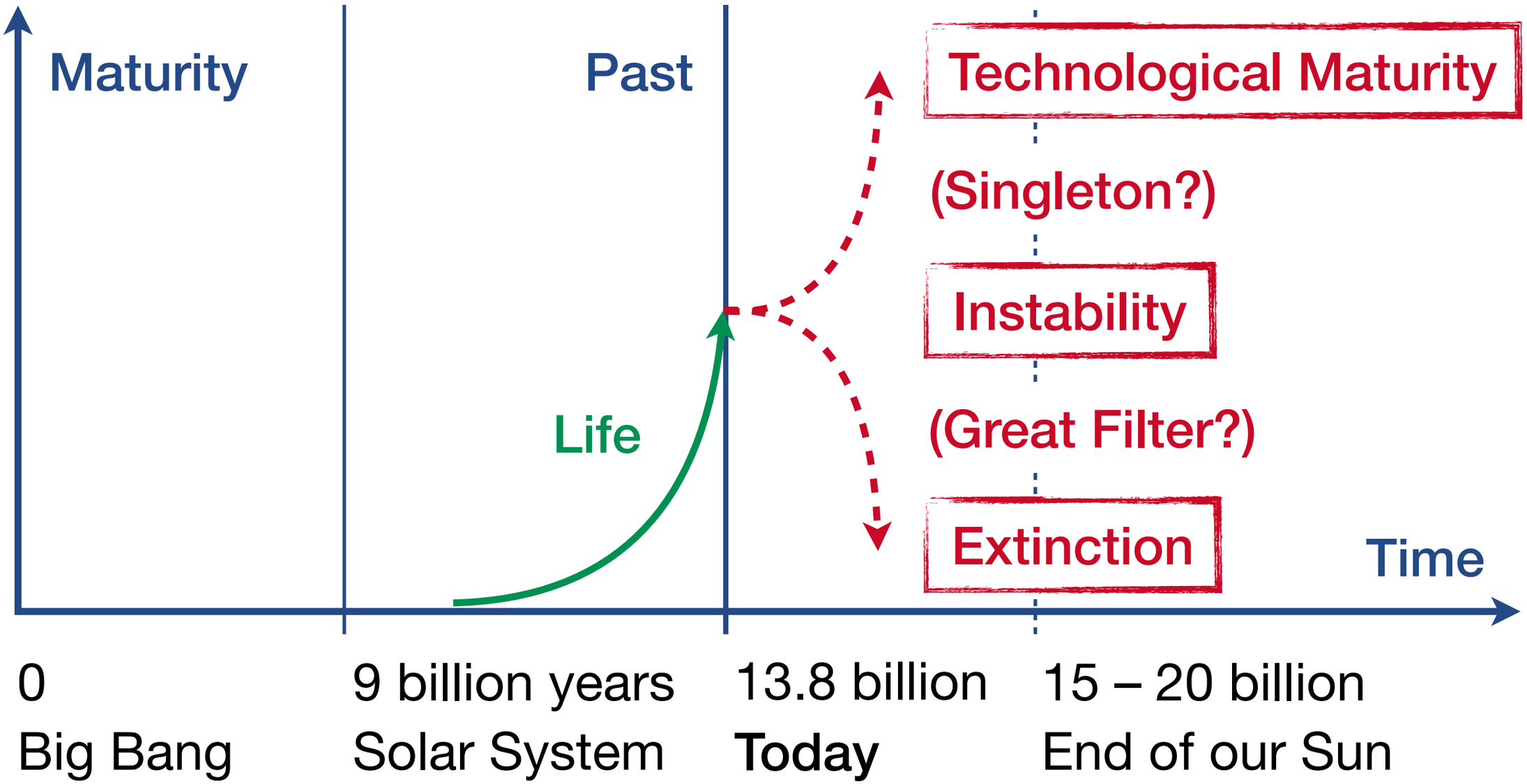
Crucial Consideration

- … an idea or argument that entails a **major change** of direction or priority.
- Overlooking just **one consideration**, our best efforts might be for naught.
- When headed the **wrong way**, the last thing we need is **progress**.



Edge: What will change everything?
edge.org/response-detail/10228

Attractor States



Bostrom: The Future of Human Evolution
www.nickbostrom.com/fut/evolution.html

Singleton … ultimate fate?

- World order with a single decision-making agency at the highest level
- Ability to prevent existential threats

Advantages:

It would avoid

- arms races
- Darwinism

Disadvantages:

It might result in a

- dystopian world
- durable lock-in

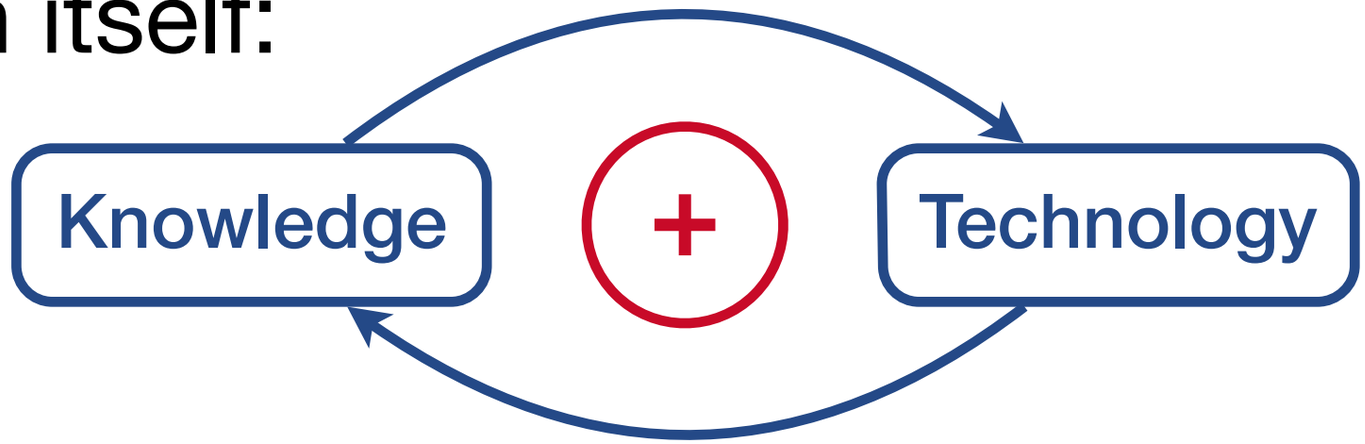
Nick Bostrom: What is a Singleton?
www.nickbostrom.com/fut/singleton.html

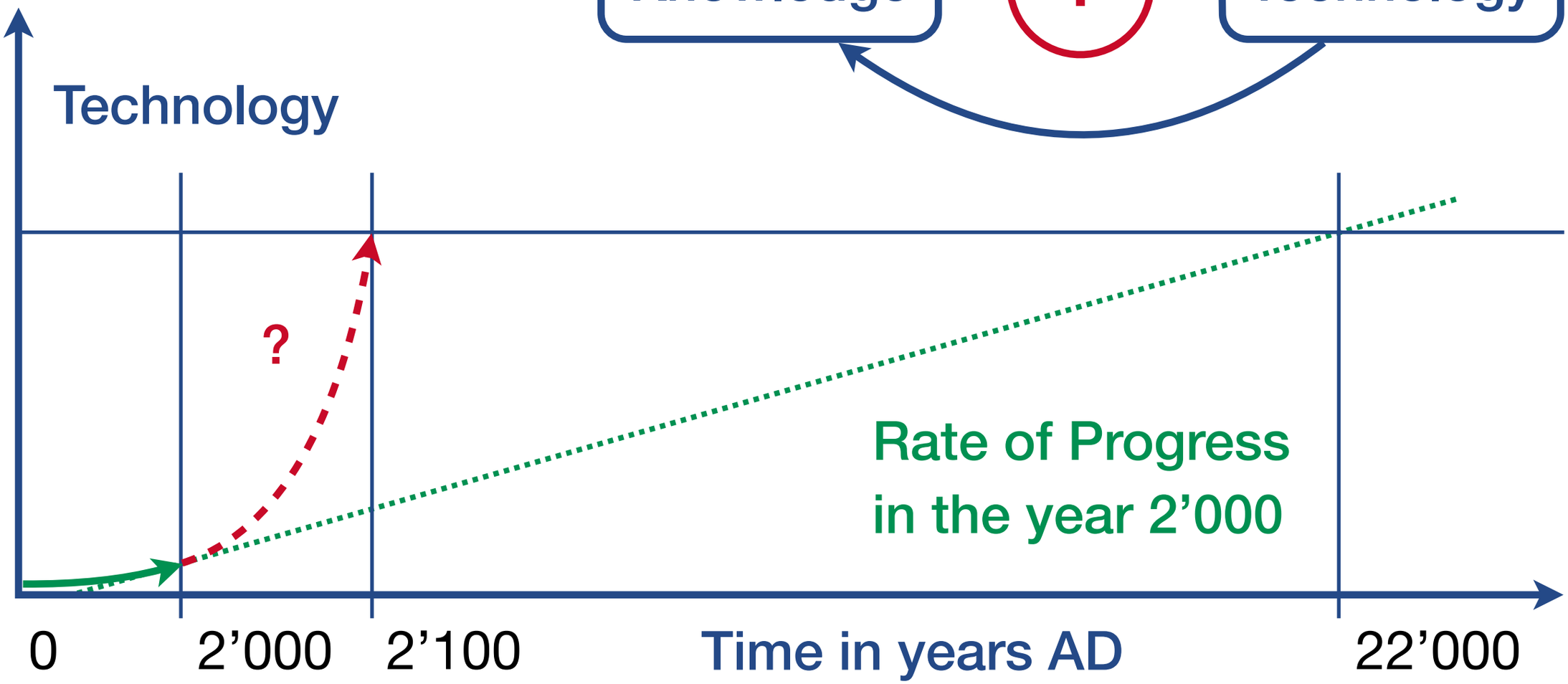
Singularity

What is the basic argument?

Accelerating Change

Progress feeds on itself:

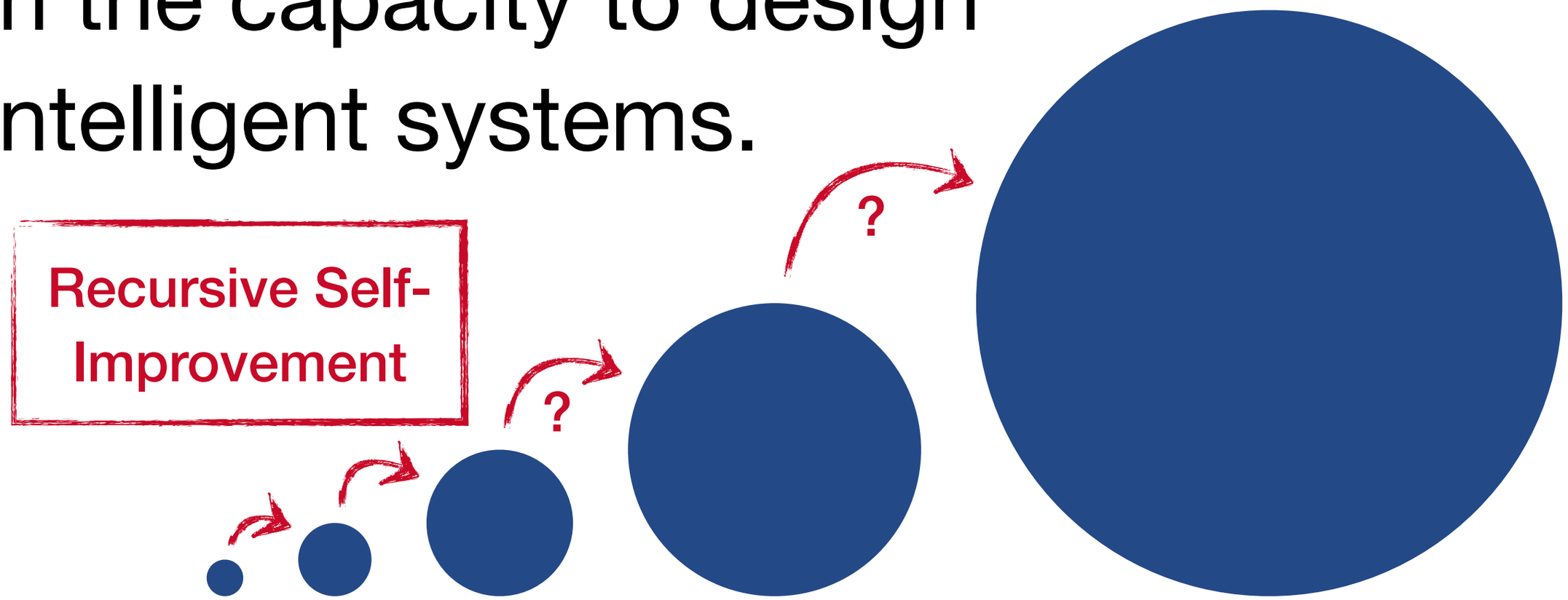




The Law of Accelerating Returns
www.kurzweilai.net/the-law-of-accelerating-returns

Intelligence Explosion

Proportionality Thesis: An increase in intelligence leads to similar increases in the capacity to design intelligent systems.



Intelligence Explosion
intelligence.org/files/IE-EI.pdf

Technological Singularity

Theoretic phenomenon: There are arguments why it should exist but it has not *yet* been confirmed experimentally.

Three major singularity schools:

- **Accelerating Change** (Ray Kurzweil)
- **Intelligence Explosion** (I.J. Good)
- **Event Horizon** (Vernor Vinge)

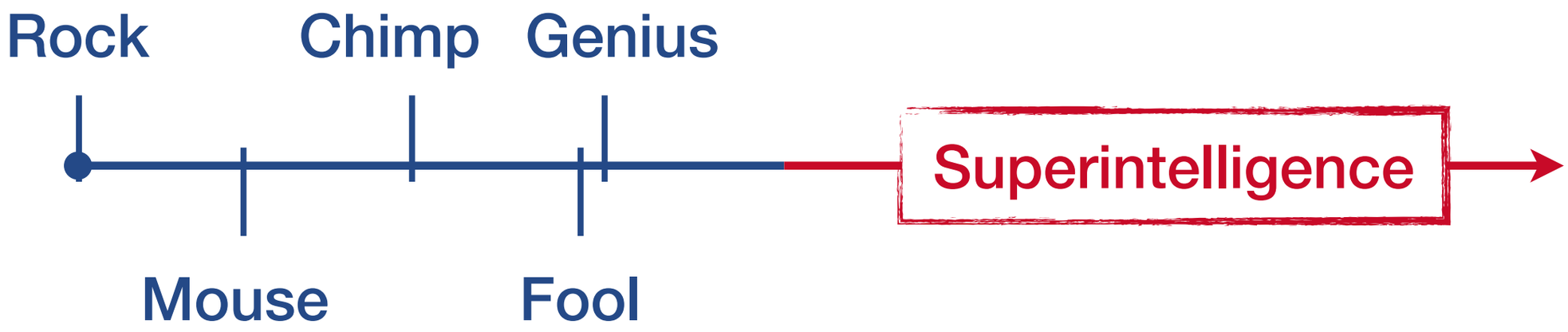
David Chalmers: The Singularity
consc.net/papers/singularity.pdf

Superintelligence

What are potential outcomes?

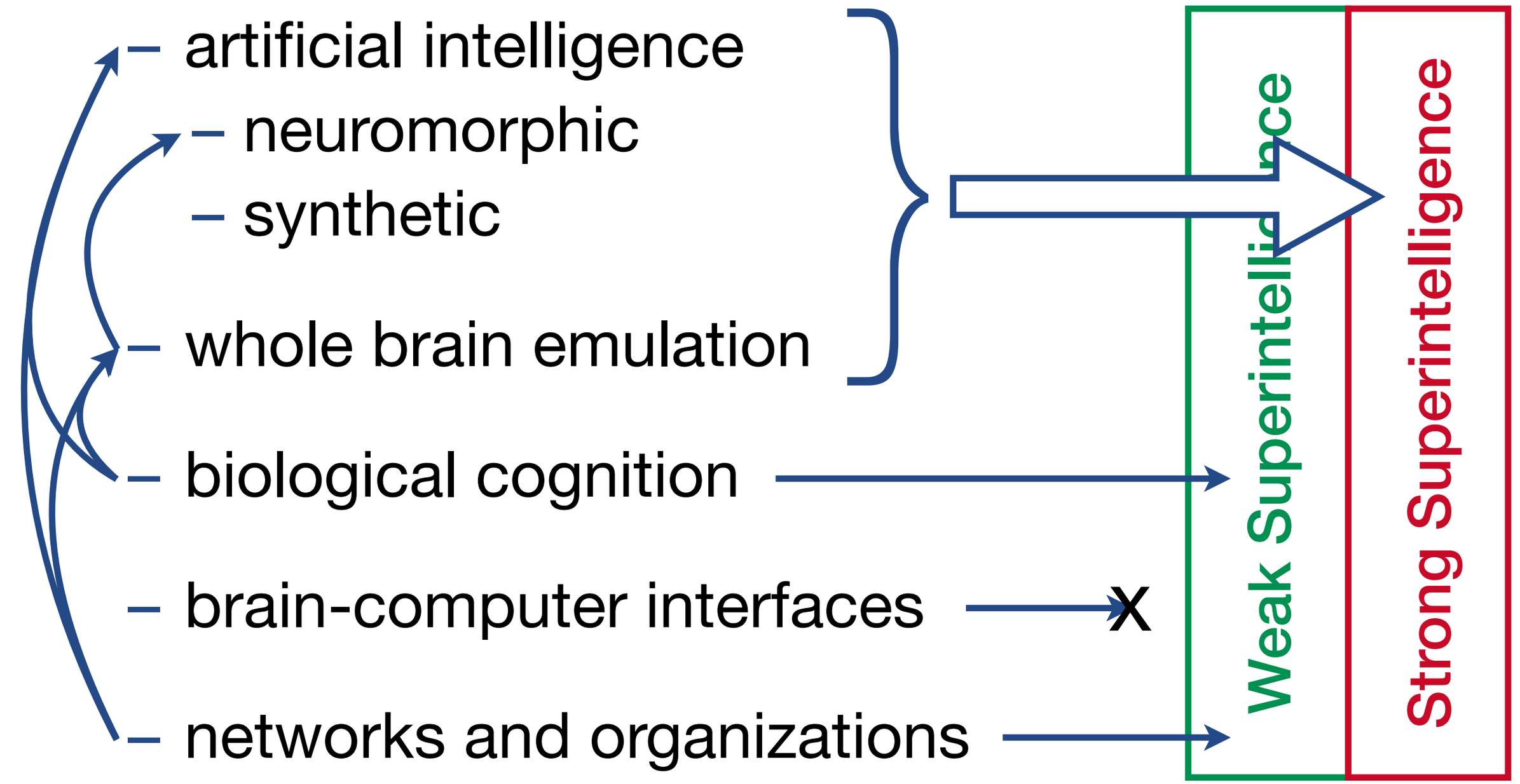
Definition of Superintelligence

An agent is called **superintelligent** if it exceeds the level of current human intelligence in all areas of interest.



Nick Bostrom: How long before Superintelligence?
www.nickbostrom.com/superintelligence.html

Pathways to Superintelligence



Embryo Selection for Cognitive Enhancement
www.nickbostrom.com/papers/embryo.pdf

Advantages of AIs over Brains

Hardware:
- Size
- Speed
- Memory

Software:
- Editability
- Copyability
- Expandability

Effectiveness:
- Rationality
- Coordination
- Communication

Human Brain	Modern Microprocessor
86 billion neurons	1.4 billion transistors
firing rate of 200 Hz	4’400’000’000 Hz
120 m/s signal speed	300’000’000 m/s

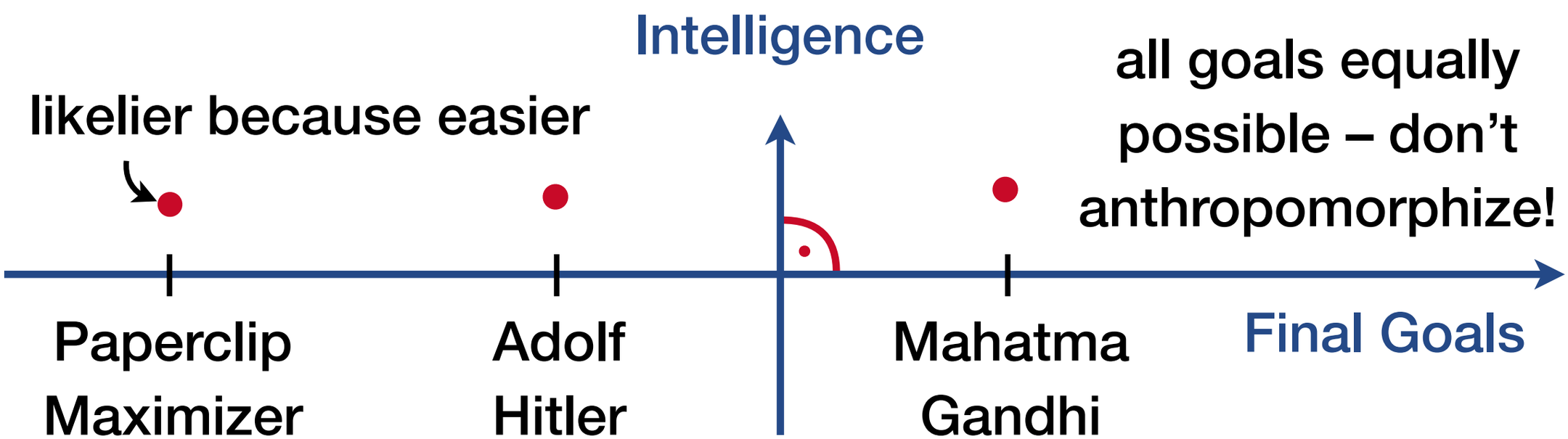
Advantages of AIs, Uploads and Digital Minds
kajsotala.fi/Papers/DigitalAdvantages.pdf

Cognitive Superpowers

- **Intelligence amplification**: bootstrapping
- **Strategizing**: overcome smart opposition
- **Hacking**: hijack computing infrastructure
- **Social manipulation**: persuading people
- **Economic productivity**: acquiring wealth
- **Technology research**: inventing new aids

Hollywood Movie Transendence
www.transcendencemovie.com

Orthogonality Thesis



Intelligence and final goals are orthogonal: Almost any level of intelligence could in principle be combined with any final goal.

Nick Bostrom: The Superintelligent Will
www.nickbostrom.com/superintelligentwill.pdf

Convergent Instrumental Goals

- **Self-Preservation**
- **Goal-Preservation**

} necessary to achieve goal

- **Resource Accumulation**
- **Intelligence Accumulation**

} to achieve goal better

Default Outcome: Doom
(Infrastructure Profusion)

Stephen M. Omohundro: The Basic AI Drives
selfawaresystems.[…].com/2008/01/ai_drives_final.pdf

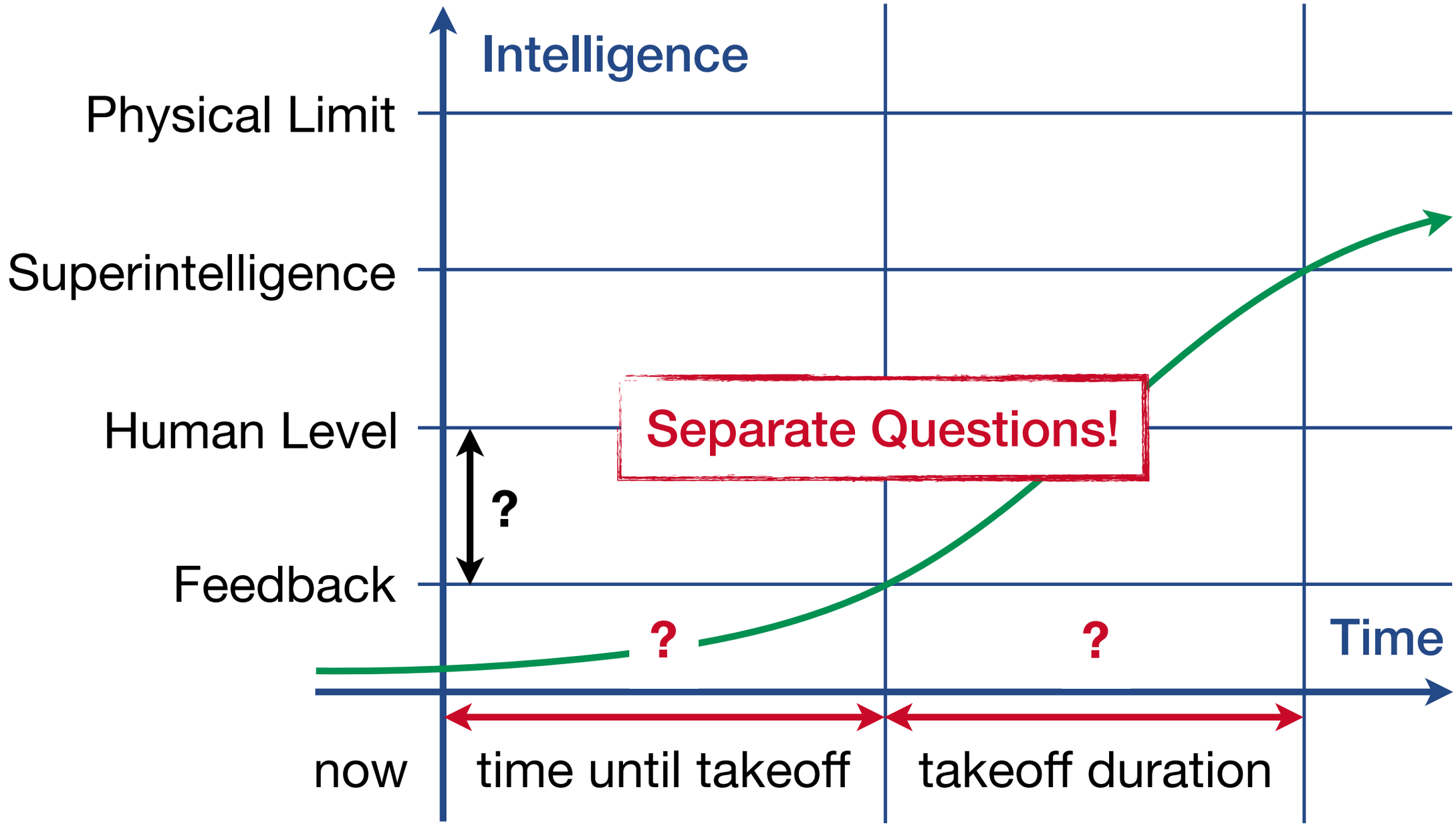
Single-Shot Situation

Our first superhuman AI must be a safe one for we may not get a second chance!

- We’re good at iterating with **testing** and **feedback**
- We’re terrible at getting things right **the first time**
- Humanity only learns when catastrophe occurred

List of Cognitive Biases
en.wikipedia.org/wiki/List_of_cognitive_biases

Takeoff Scenarios



The Hanson-Yudkowsky AI-Foom Debate
intelligence.org/files/AIFoomDebate.pdf

Potential Outcomes

Fast Takeoff	Slow Takeoff
hours, days, weeks	several months, years
Unipolar Outcome	**Multipolar Outcome**
Singleton	**Second Transition**
(Slide 10)	Unification by Treaty

State and Trends

Where are we heading to?

Brain vs. Computer

Brain		Computer	
Consciousness	sequential	**Software**	parallel
Mindware	parallel	**Hardware**	sequential

easy	**Pattern Recognition**	hard	parallel GPUs
hard	**Logic and Thinking**	easy	

… but there is massive progress*!

* We have superhuman image recognition since February 2015.

Dennett: Consciousness Explained
www.amazon.com/dp/0316180661

State of the Art

Deep Blue: 1997

Stanley: 2005

IBM Watson: 2011

Schmidhuber: 2011

Checkers	Superhuman
Backgammon	Superhuman
Othello	Superhuman
Chess	Superhuman
Crosswords	Expert Level
Scrabble	Superhuman
Bridge	Equal to Best
Jeopardy!	Superhuman
Poker	Varied
FreeCell	Superhuman
Go	Strong Amateur

How bio-inspired deep learning keeps winning competitions
www.kurzweilai.net/how-bio-inspired-deep-learning-[…]

Machine Learning by Google

Vicarious AI passes first Turing Test: CAPTCHA
news.vicarious.com/[…]-ai-passes-first-turing-test

Predicting AI Timelines

Great uncertainties:

- Hardware or software the bottleneck?
- Small team or a Manhattan Project?
- More speed bumps or accelerators?

Probability for AGI	10%	50%	90%
AI scientists, median	2024	2050	2070
Luke Muelhauser, MIRI	2030	2070	2140

How We're Prediciting AI – or Failing To
intelligence.org/files/PredictingAI.pdf

Speed Bumps

- Depletion of low-hanging fruit
- An end to Moore's law
- Societal collapse
- Disinclination

Accelerators

- Faster **hardware**
- Better **algorithms**
- Massive **datasets**

+ enormous economic, military and egoistic incentives!

Machine Intelligence Research Institute: When AI? intelligence.org/2013/05/15/when-will-ai-be-created/

Strategy

What is to be done?

Prioritization

- **Scope**: How big/important is the issue?
- **Tractability**: What can be done about it?
- **Crowdedness**: Who else is working on it?

Work on the matters that matter the most!

- AI is **the key lever** on the long-term future
- Issue is **urgent**, **tractable** and **uncrowded**
- The stakes are **astronomical**: our light cone

Luke Muehlhauser: Why MIRI?
intelligence.org/2014/04/20/why-miri/

Flow-Through Effects

Going meta: Solve the problem-solving problem!

Holden Karnofsky: Flow-Through Effects
blog.givewell.org/2013/05/15/flow-through-effects/

Controlled Detonation

Difficulty:

Friendly AI >> General AI

AI as a Positive and Negative Factor in Global Risk
intelligence.org/files/AIPosNegFactor.pdf

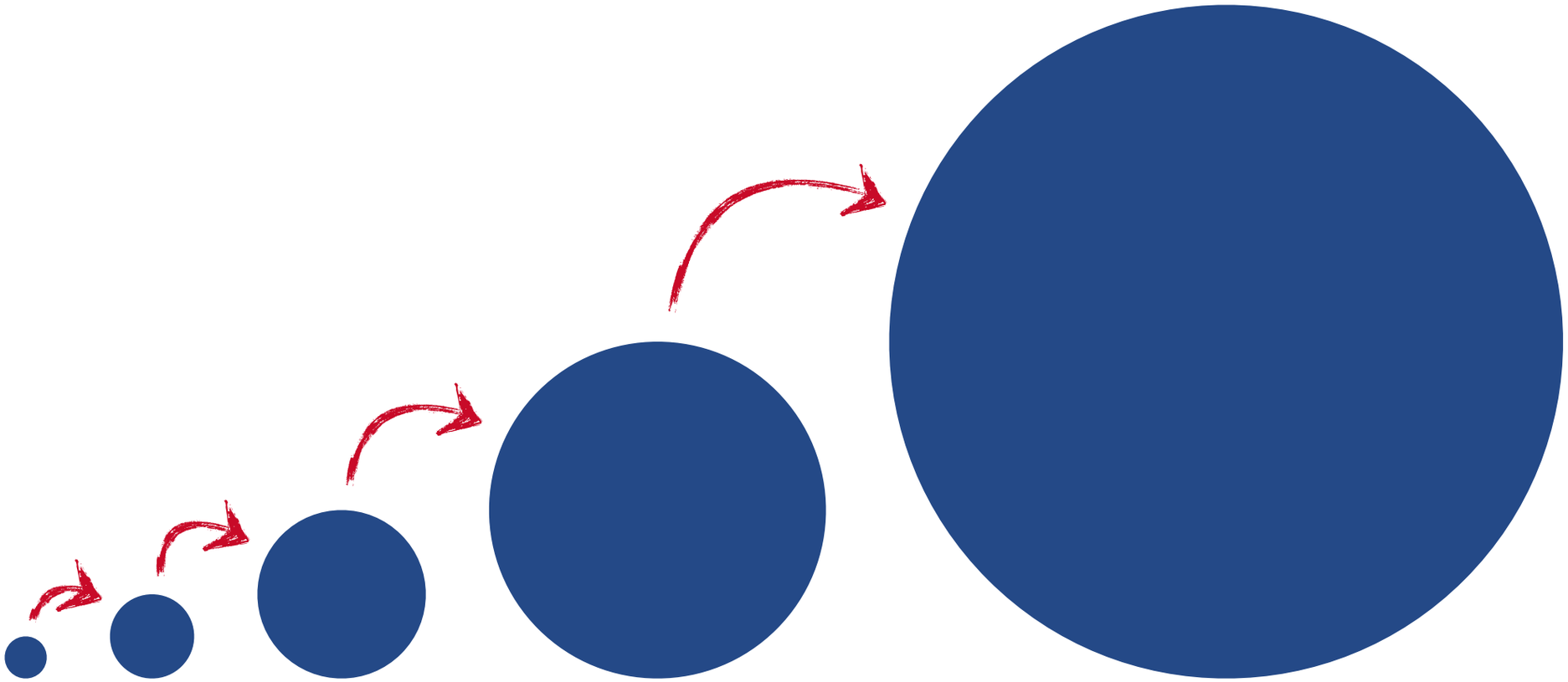
Control Problem

Will AI outsmart us?

Capability Control	Motivation Selection
Boxing	Direct Specification
Stunting	Indirect Normativity
Tripwires	Incentive Methods

Roman V. Yampolskiy: Leakproofing the Singularity
cecs.louisville.edu/ry/LeakproofingtheSingularity.pdf

Stable Self-Improvement



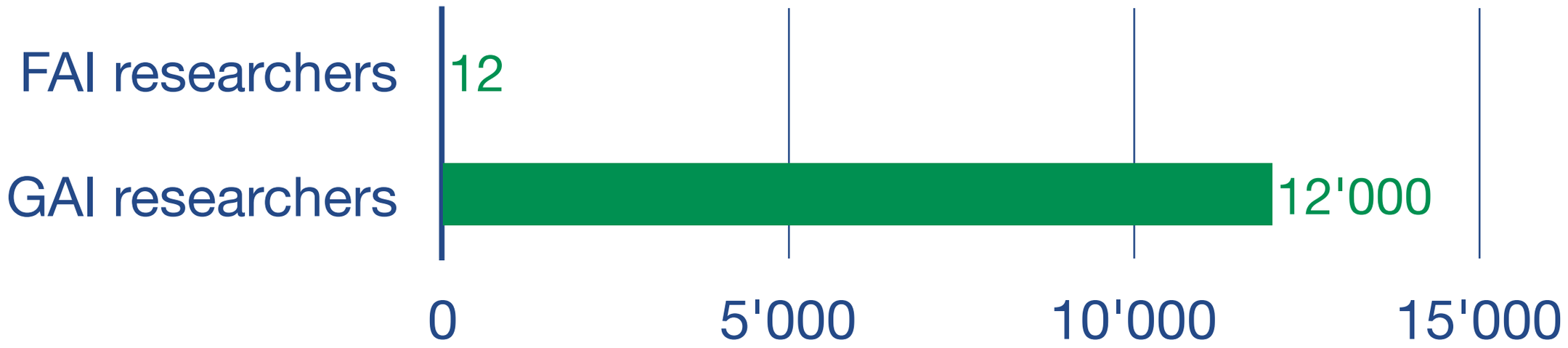
Friendly

Friendly?

Differential Intellectual Progress

Prioritize risk-reducing intellectual progress over risk-increasing intellectual progress

AI safety should outpace **AI capability** research



International Cooperation

- We are the ones who will create superintelligent AI
- Not primarily a **technical problem**, rather a social
- **International regulation?**

In face of uncertainty, cooperation is robust!

Summary

What have we learned?

Crucial Crossroad

**Instead of passively drifting,
we need to steer a course!**

- **Philosophy**
- **Mathematics**
- **Cooperation**

with a deadline.

Luke Muehlhauser: Steering the Future of AI
intelligence.org/[…]Steering-the-Future-of-AI.pdf

«Before the prospect of an intelligence explosion, we humans are like children playing with a bomb. Such is the mismatch between the power of our play-thing and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound.»

— Prof. Nick Bostrom in his book *Superintelligence*

gbs switzerland

Discussion

www.superintelligence.ch

Kaspar Etter, kaspar.etter@gbs-schweiz.org
Adrian Hutter, adrian.hutter@gbs-schweiz.org