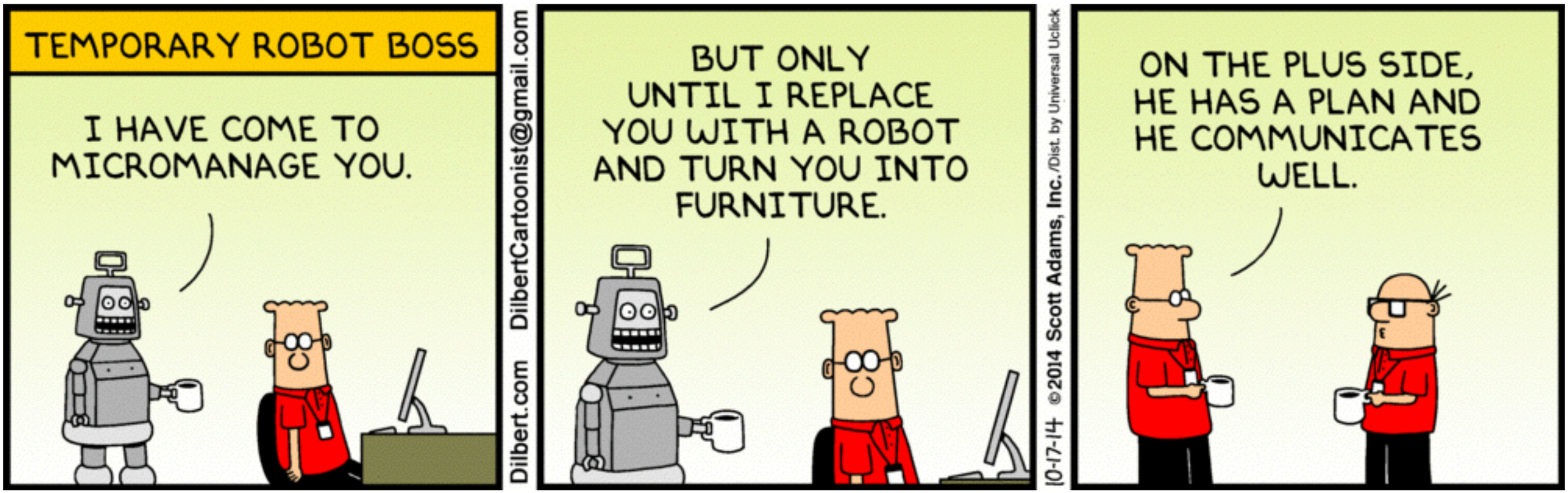
gbs giordano bruno stiftung **schweiz**

Superintelligence

Our Final Invention

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

Basel, Switzerland
22 November 2014



«Artificial Intelligence makes philosophy honest.»
— Daniel Dennett (2006), American Philosopher

Outline

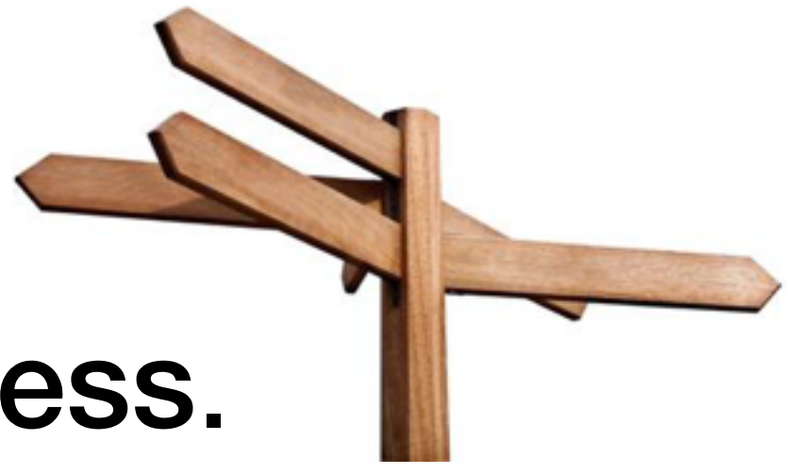
- Introduction
- Singularity
- Superintelligence
- State and Trends
- Strategy
- Sources
- Summary

Introduction

What are we talking about?

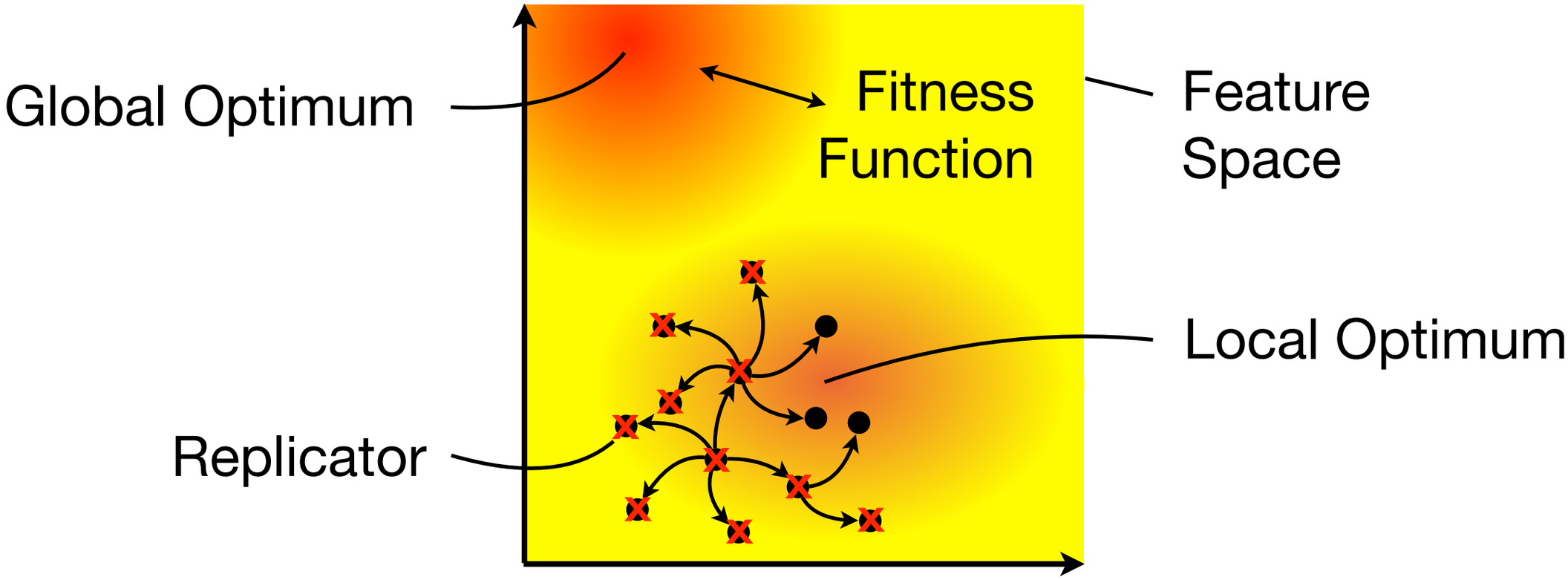
Crucial Consideration

- … 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

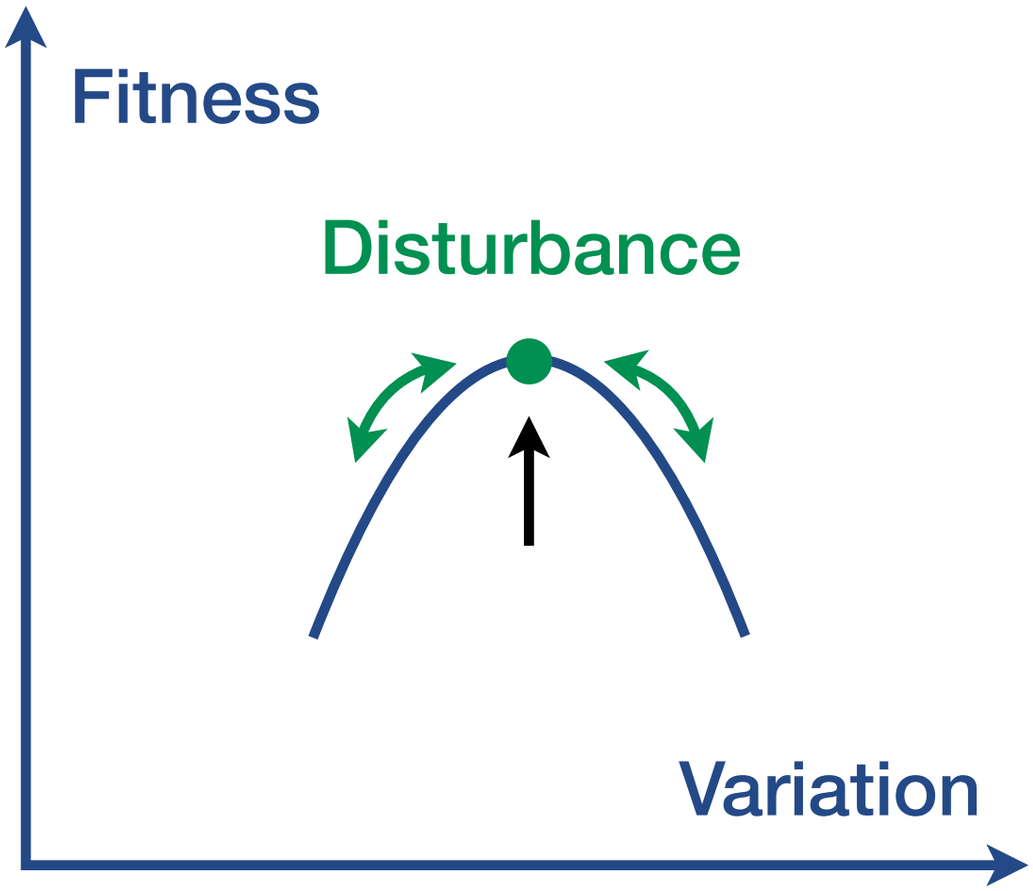
Evolution



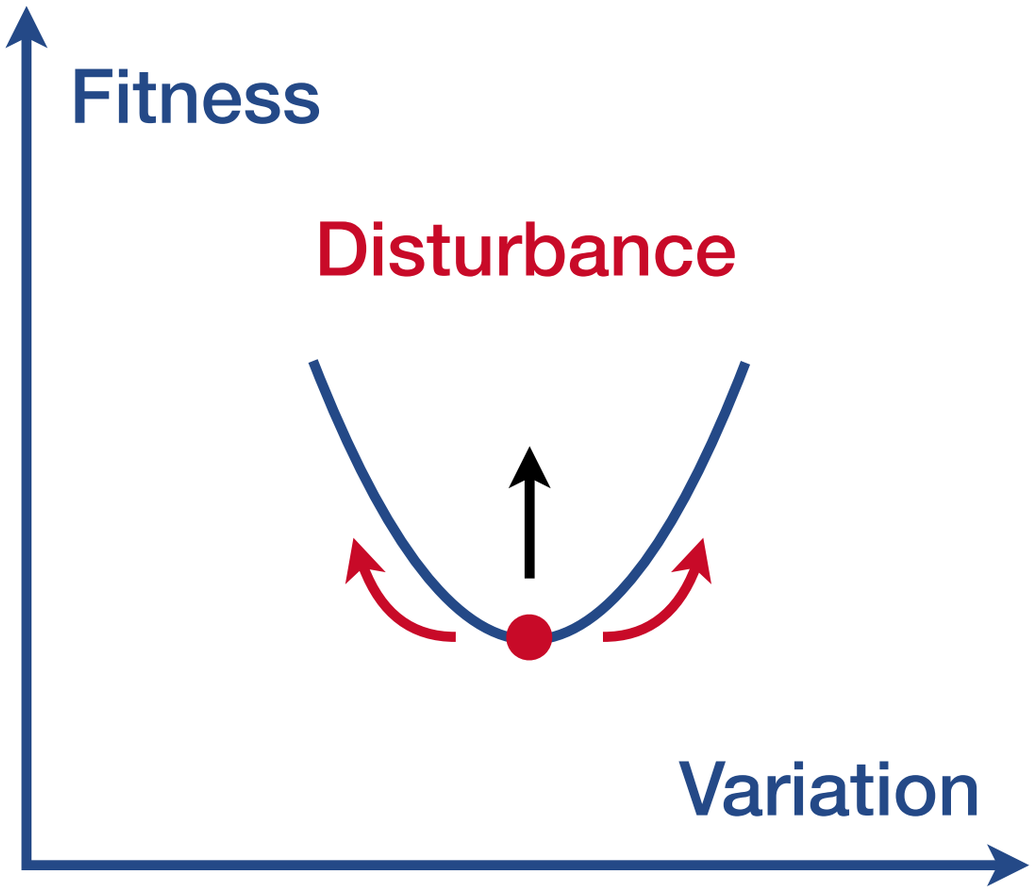
Foresight is the power of intelligence!

Hill Climbing Algorithm & Artificial Intelligence
www.youtube.com/watch?v=oSdPmxRCWws

Stability



stable states stay
(when time passes)

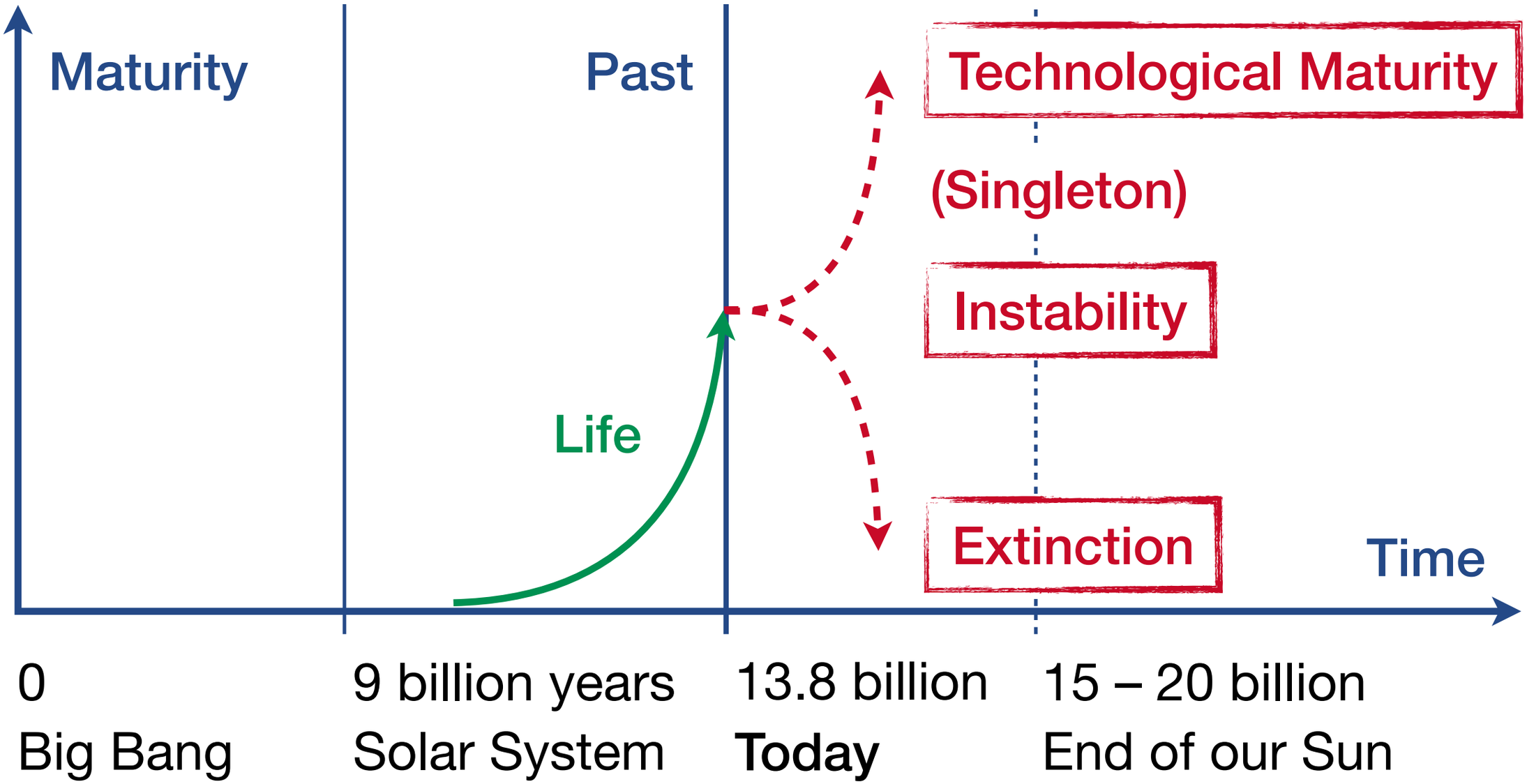


instable states vanish
(unless they are cyclic)

Richard Dawkins: The Selfish Gene
www.amazon.com/dp/0199291152/

Attractors

Big Rip: ≥ 20 billion years from now
Big Crunch: ≥ 10^2 billion years from now
Big Freeze: ≥ 10^5 billion years from now
Heat Death: ~ 10^{1000} years from now



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:

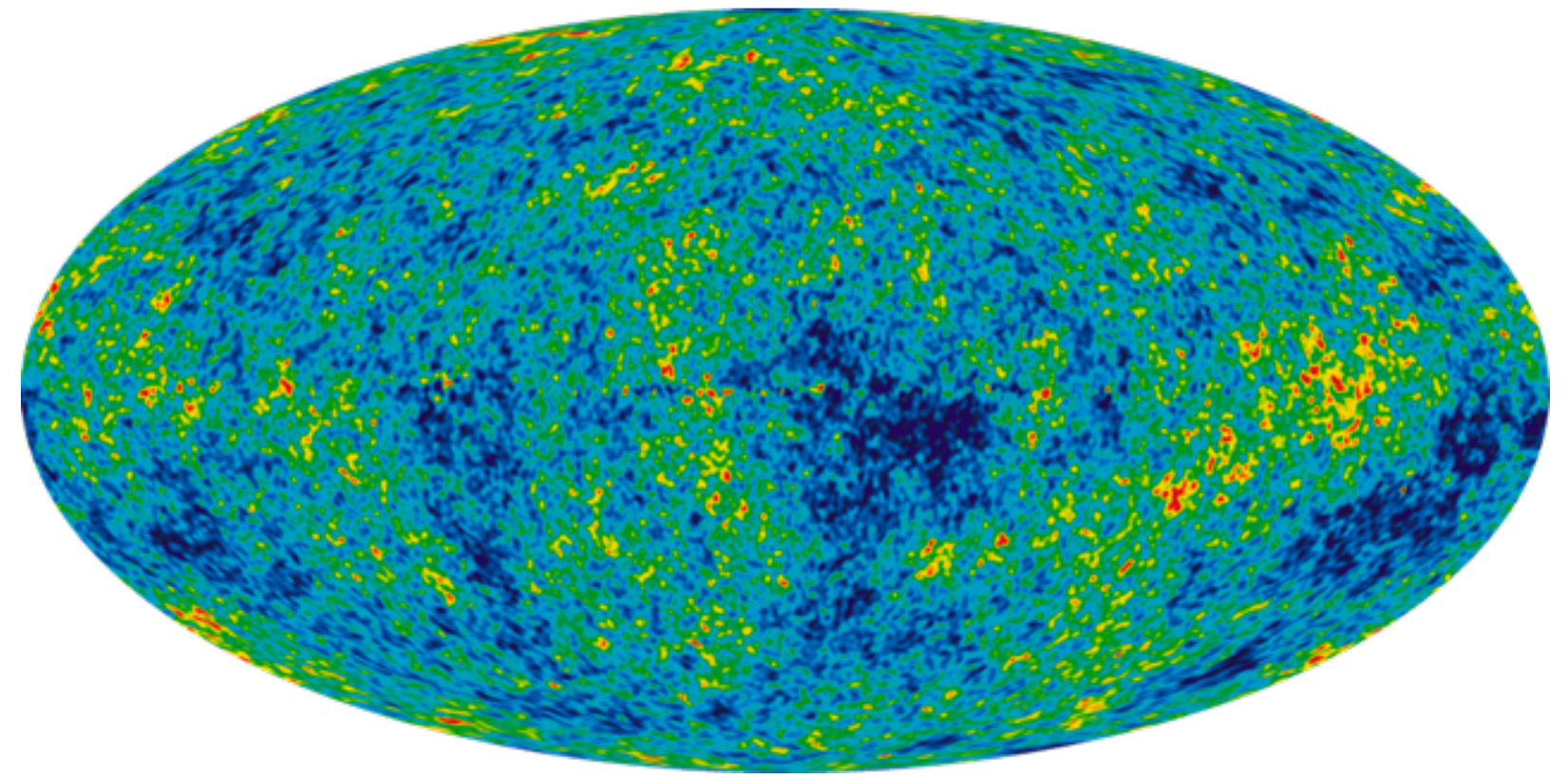
It might result in a

- dystopian world
- durable lock-in

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

The (Observable) Universe

93 billion light years



$> 10^{11}$ galaxies

100000000000

$\sim 3 \cdot 10^{23}$ stars

300000000000000000000000

Universe
en.wikipedia.org/wiki/Universe

Fermi Paradox

Where are they? (Extraterrestrial Life)

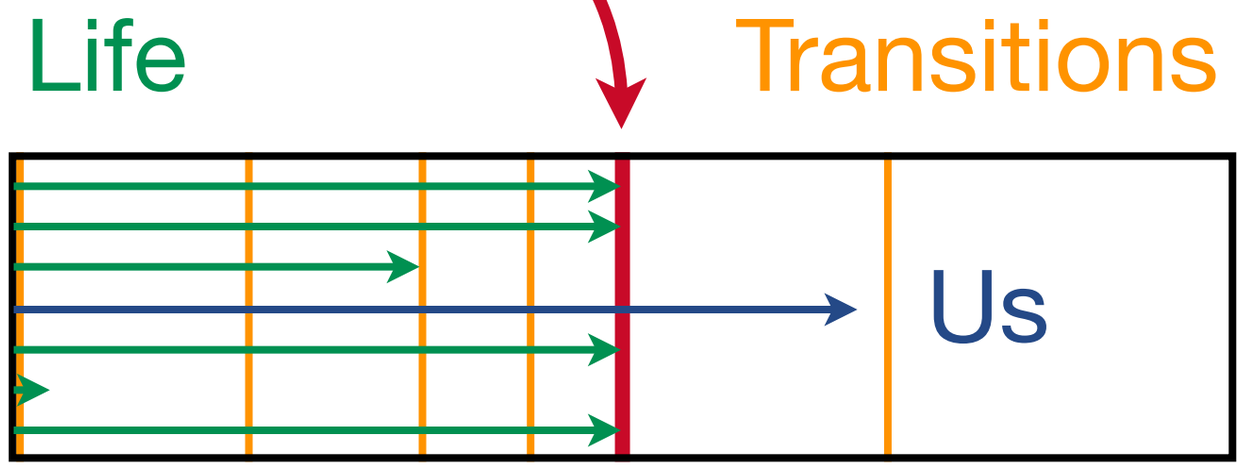
There are two groups of explanation:

- **There are none**, i.e. we're all alone.
- **We can't detect them** because …
  - we're too primitive or too far apart
  - there are predators or all fear them
  - we're lied to, live in a simulation, …

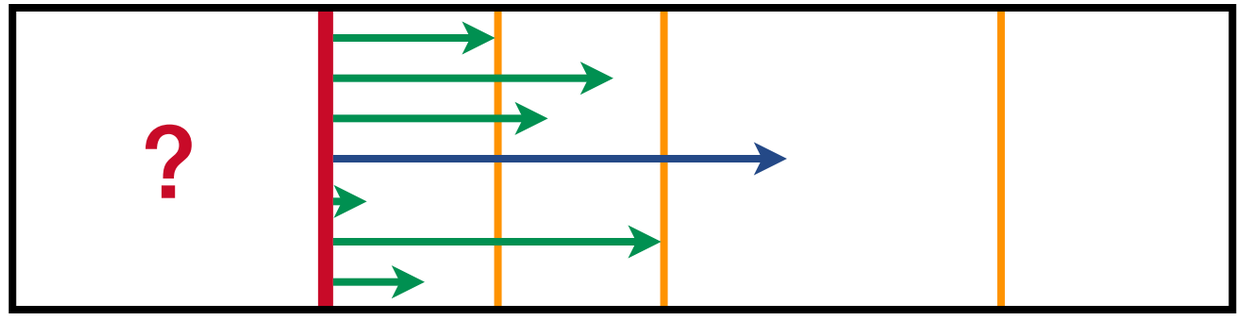
Fermi Paradox
en.wikipedia.org/wiki/Fermi_paradox

Great Filter

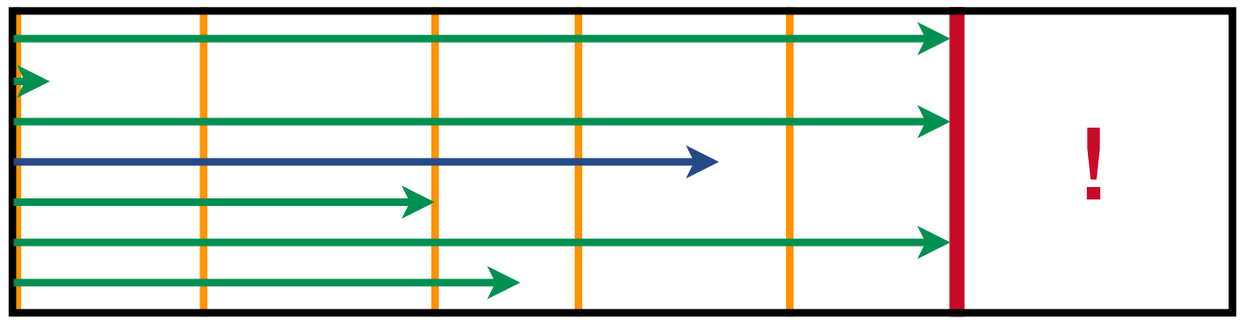
We're rare:



We're the first:

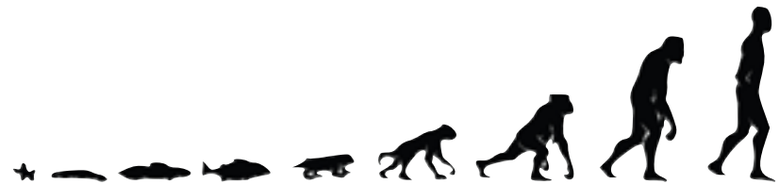


We're doomed:

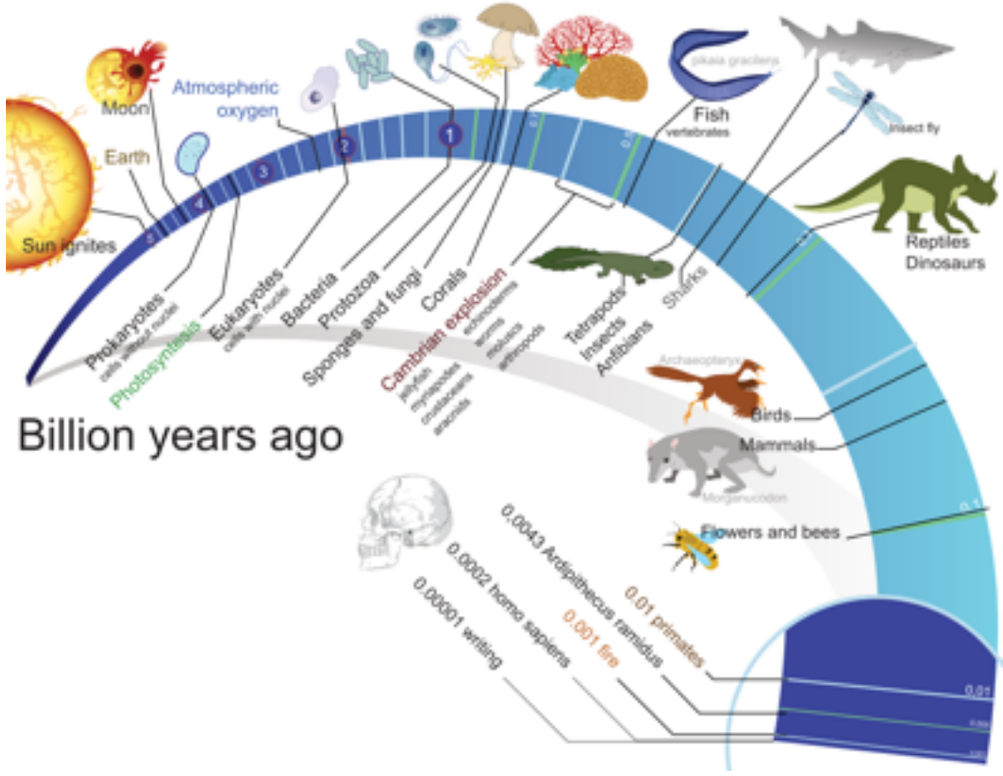


The Fermi Paradox
waitbutwhy.com/2014/05/fermi-paradox.html

Major Transitions



- Self-replicating molecules (**abiogenesis**)
- Simple (**prokaryotic**) single-cell life
- Complex (**eukaryotic**) single-cell life
- Sexual reproduction
- Multi-cell organisms
- Tool-using animals
- Where we are now
- Space colonization



The Major Transitions in Evolution
www.amazon.com/dp/019850294X/

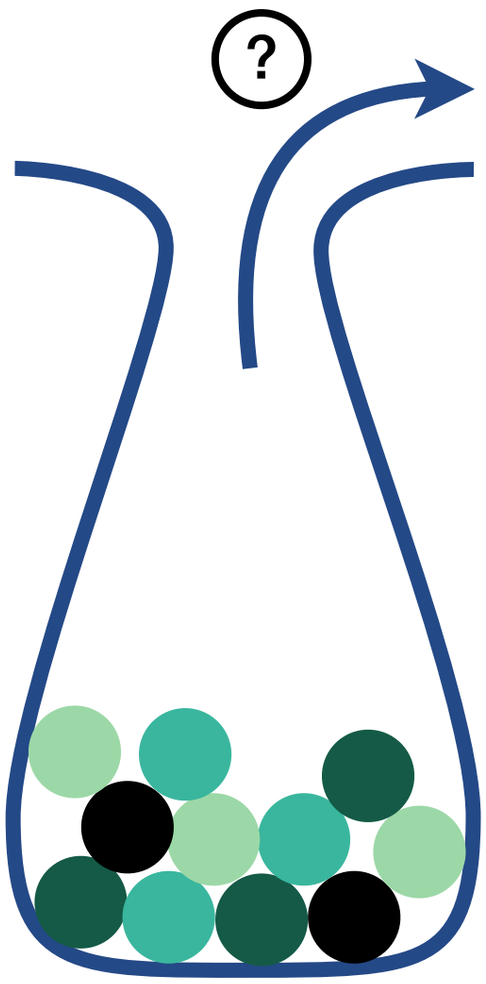
Anthropic Principle

- How probable are these transitions?
- They have occurred **at least once**
- Observation is conditional on existence

$$P(\text{complex life on earth} \mid \text{our existence}) = 1$$

There are observer selection effects!

Technologies



- Taking balls out of a jar
- No way to put back in
- **Black balls** are lethal

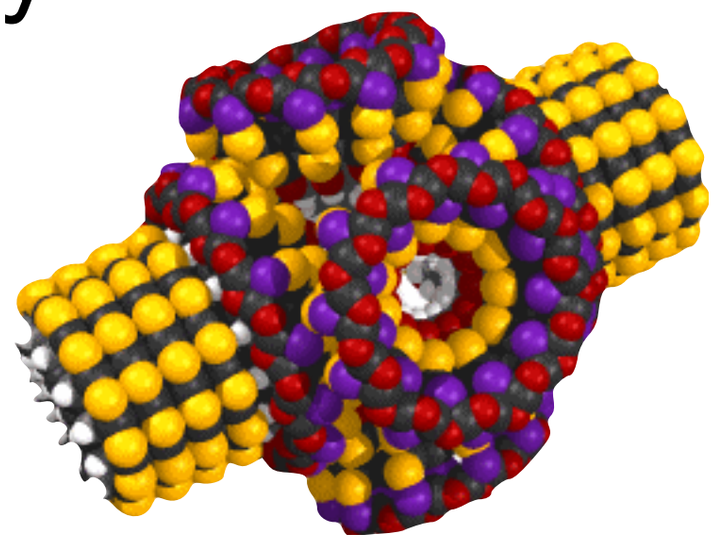
By definition:

- No ball has been black
- We'll only take out one

Nick Bostrom @ Google
youtu.be/pywF6Zzsghl?t=9m

Candidates

- Nuclear Weapons (still possible)
- Synthetic Biology (engineered pathogens)
- Totalitarianism enabling Technologies
- Molecular Nanotechnology
- Machine Intelligence
- Geoengineering
- **Unknown**



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:

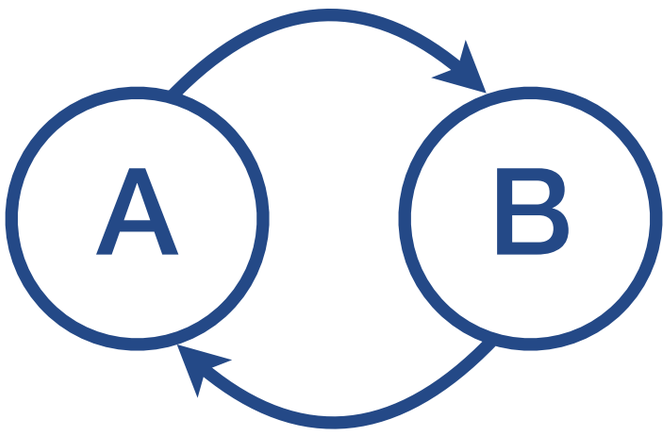
- Machines deserve moral consideration
- We might live in a computer simulation

Singularity

What is the basic argument?

Feedback

Systems can feed back into themselves and thus must be **analyzed as a whole**!



Feedback is either:

- **Positive** (reinforcing)
- **Negative** (balancing)

Feedback
en.wikipedia.org/wiki/Feedback

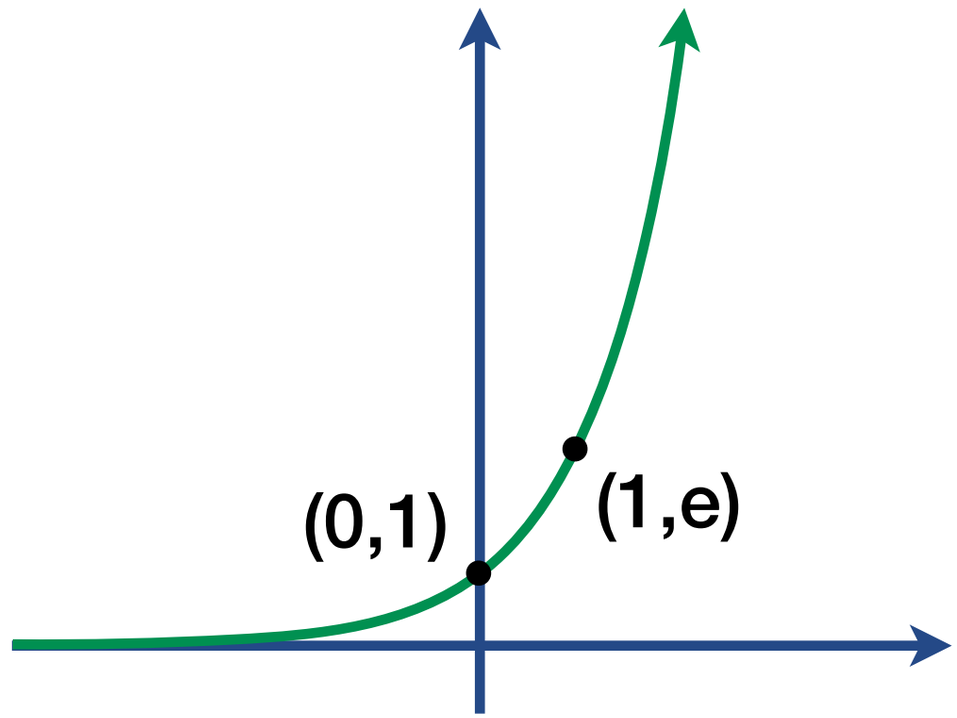
Exponential Functions

If increase is linear to current amount:

$$\frac{d}{dx} f(x) = c \cdot f(x)$$

solved by

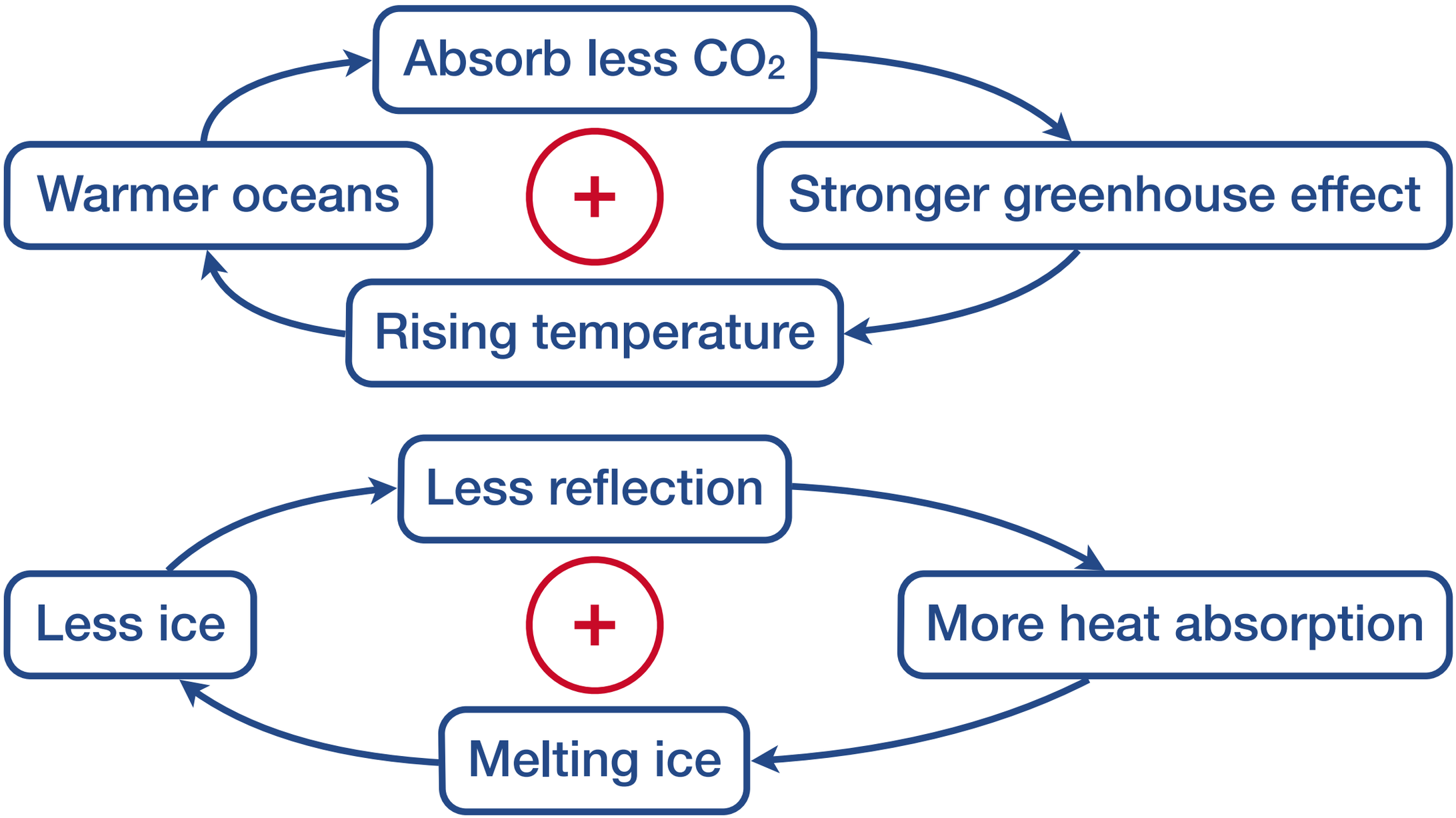
$$f(x) = e^{c \cdot x}$$



Fold a paper 45 times $\Longrightarrow$ to the moon!

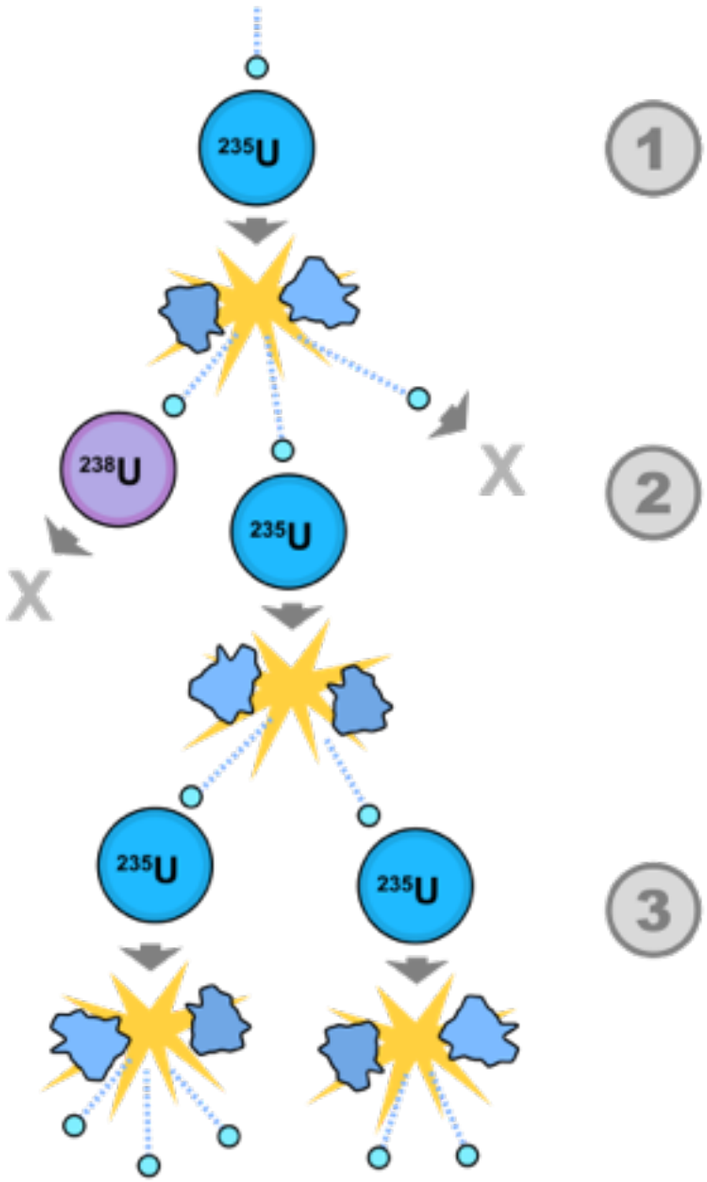
How folding a paper can get you to the moon
www.youtube.com/watch?v=AmFMJC45f1Q

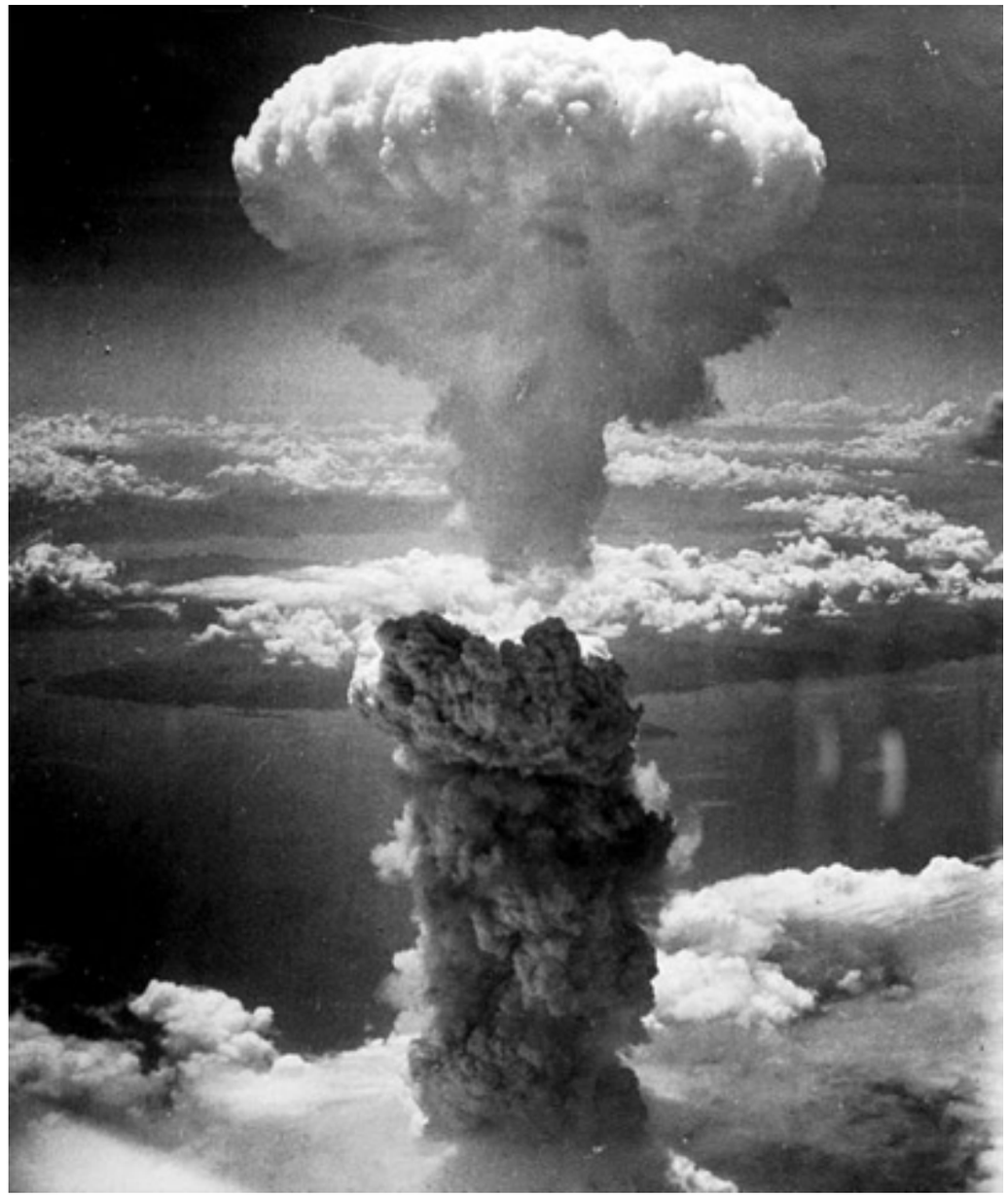
Climate Change



Climate Change Feedback
en.wikipedia.org/wiki/Climate_change_feedback

Nuclear Chain Reaction

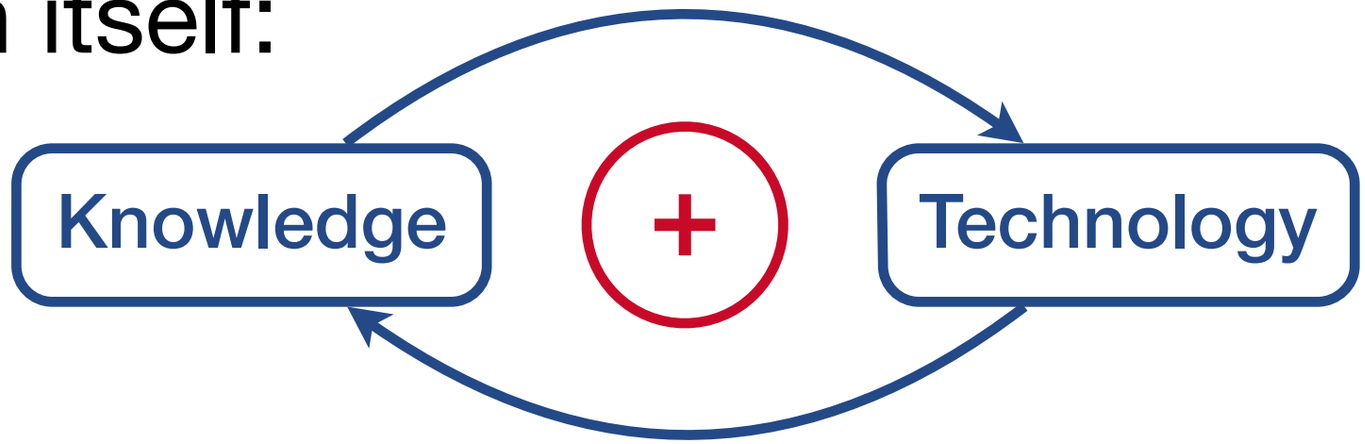


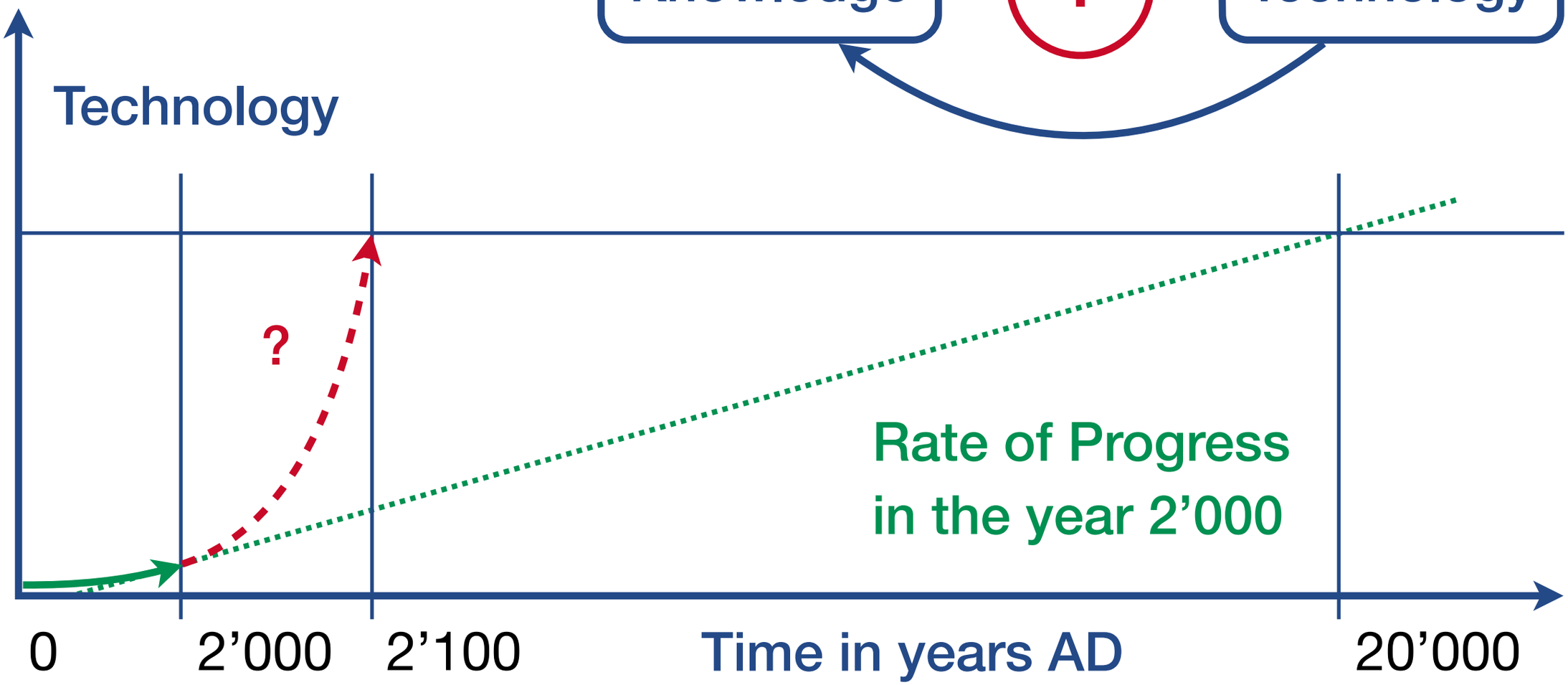


Nuclear Chain Reaction
en.wikipedia.org/wiki/Nuclear_chain_reaction

Accelerating Change

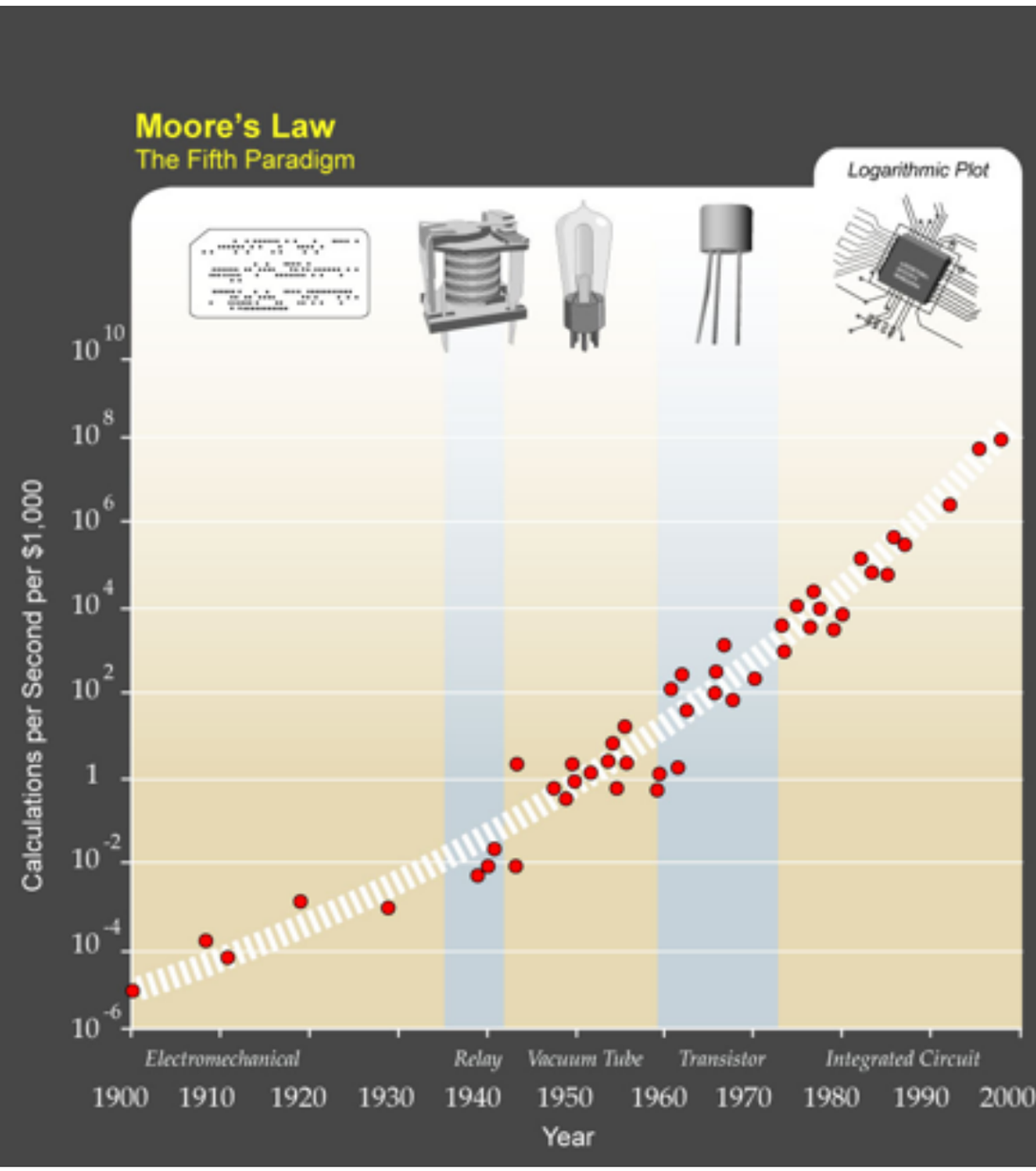
Progress feeds on itself:

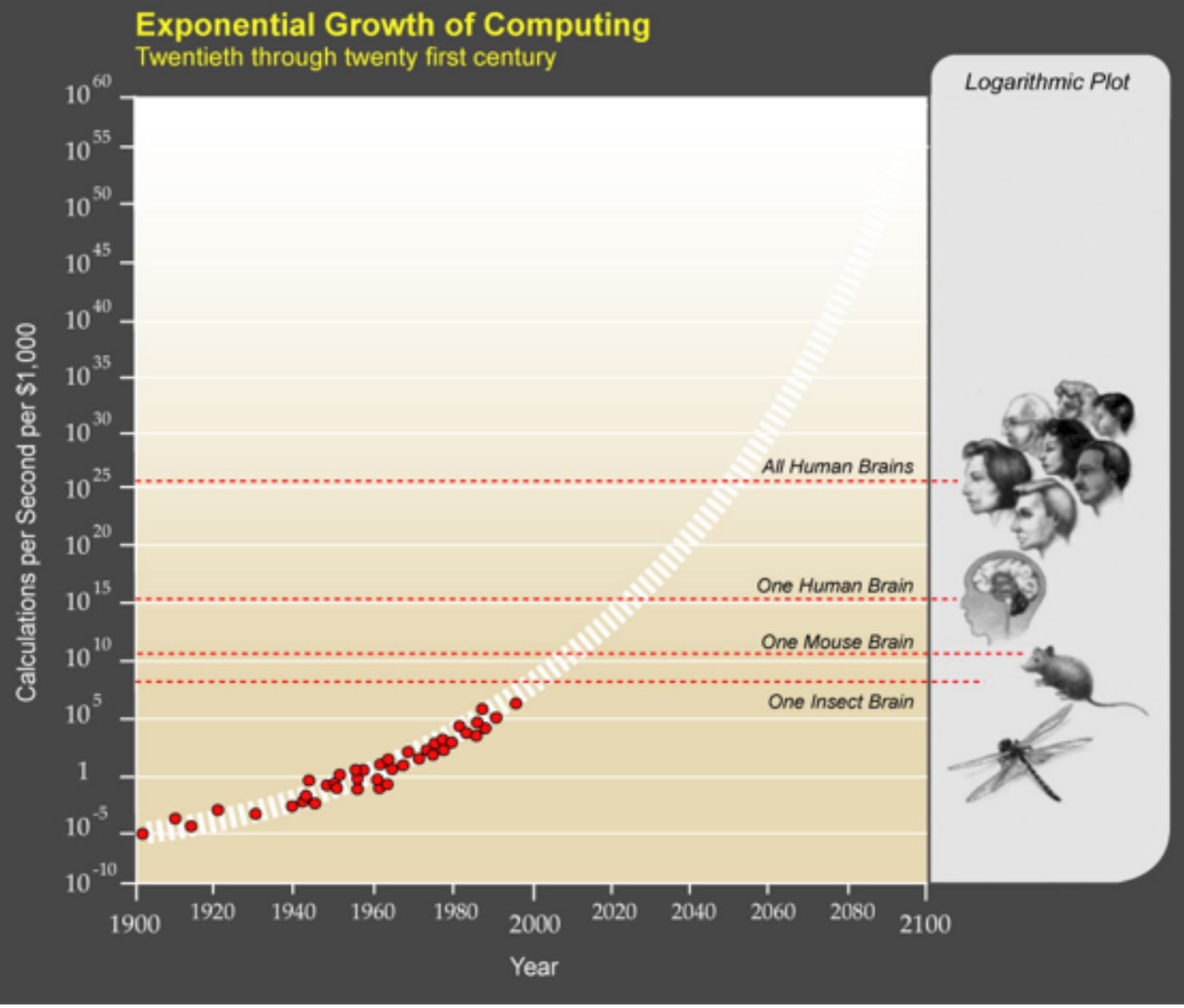




The Law of Accelerating Returns
www.kurzweilai.net/the-law-of-accelerating-returns

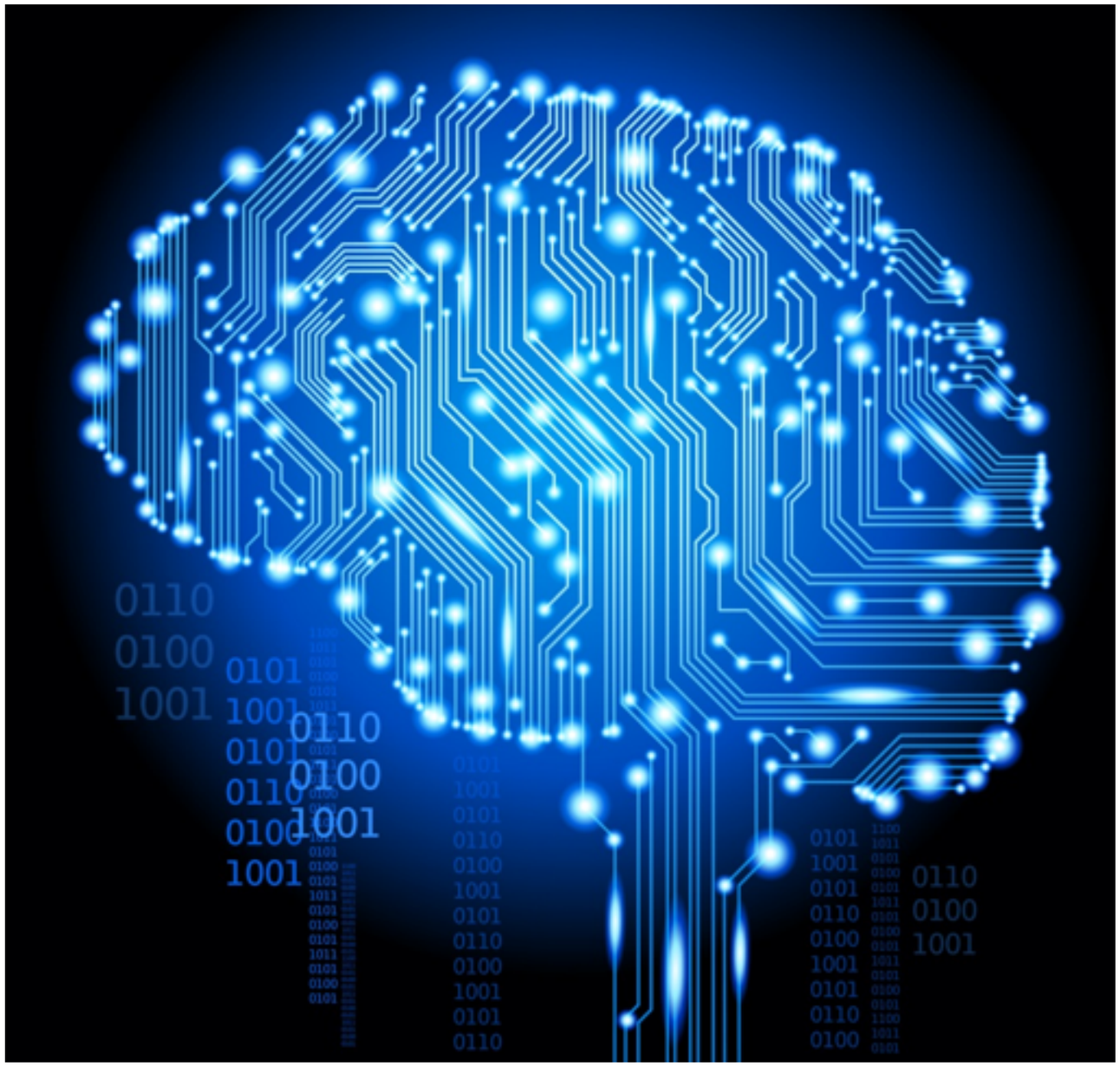
Moore’s Law

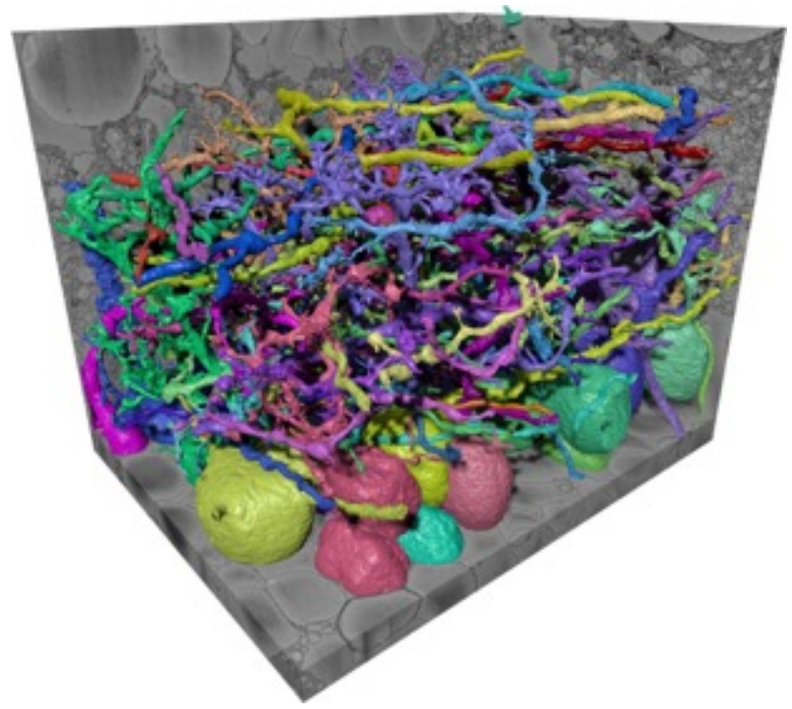




Exponential and Non-Exponential Trends in IT
intelligence.org/[…]/exponential-and-non-exponential/

Artificial Mind





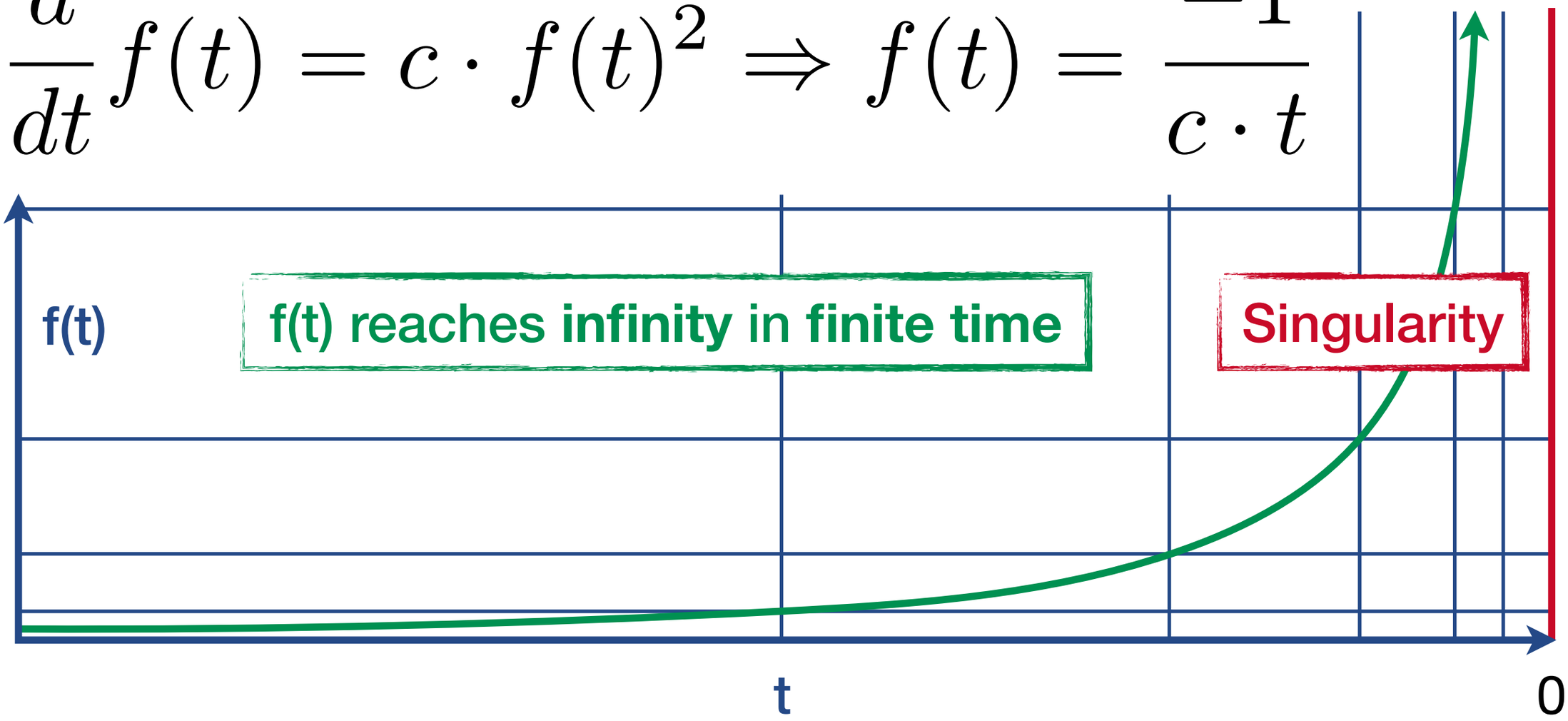
Imagine all relevant aspects captured in a **computer model** (thought experiment)

Whole Brain Emulation: A Roadmap
www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf

Hyperbolic Growth

Second-order positive feedback loop

$$\frac{d}{dt} f(t) = c \cdot f(t)^2 \Rightarrow f(t) = \frac{-1}{c \cdot t}$$



Mathematical Singularity
en.wikipedia.org/wiki/Singularity_(mathematics)

Speed Explosion

Time

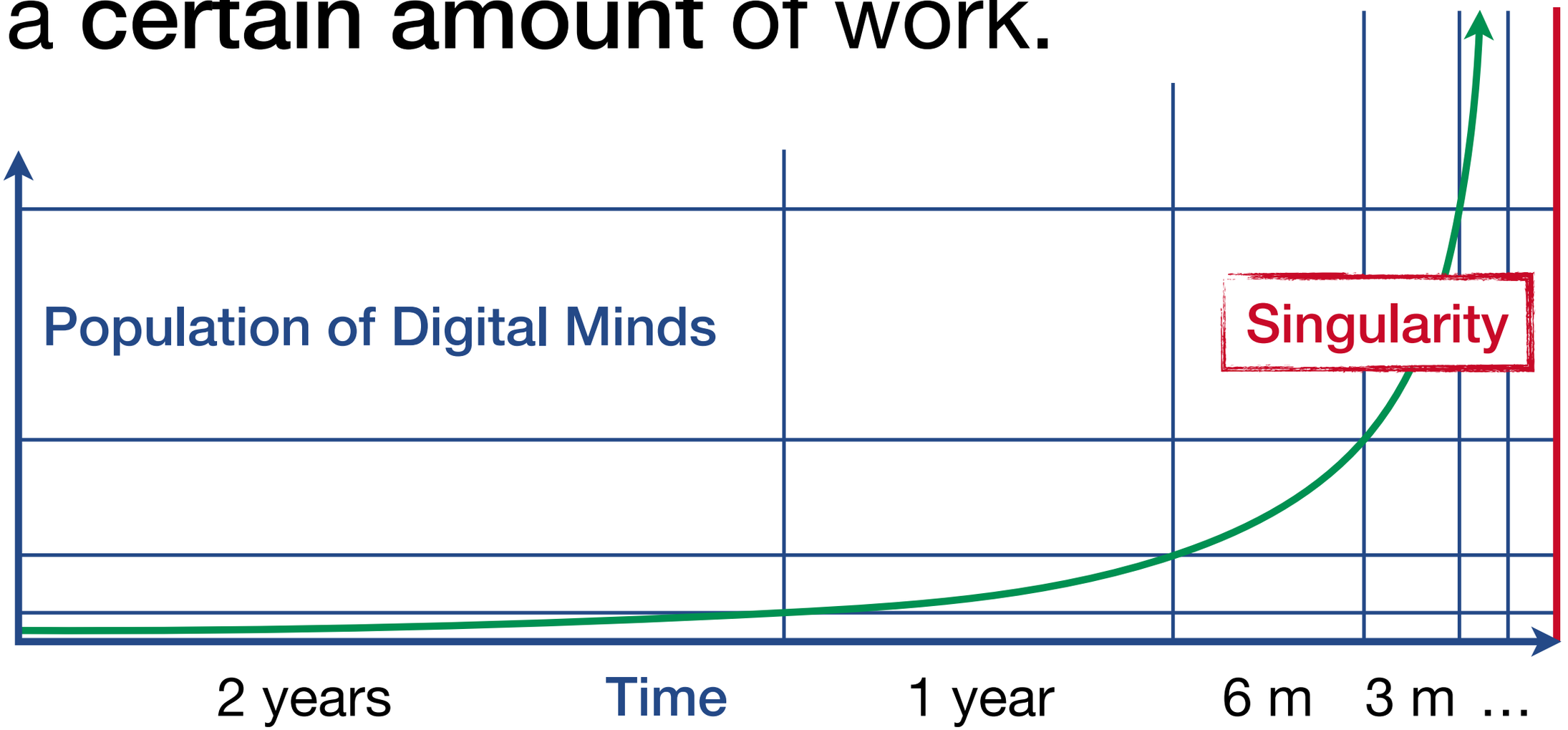
Computing speed doubles every two **subjective years** of work.

Marcus Hutter: Can Intelligence Explode?
www.hutter1.net/publ/singularity.pdf

Population Explosion

Quantitative

Computing costs halve for
a **certain amount** of work.

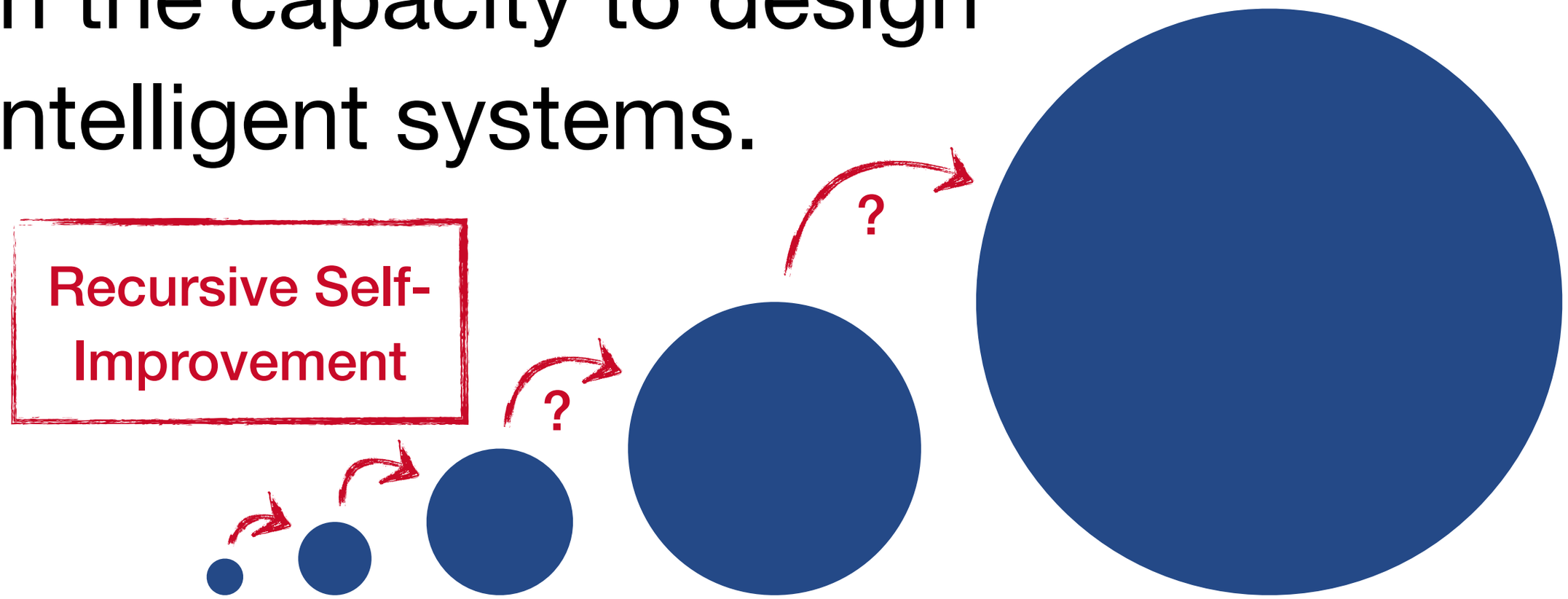


Ray Solomonoff: The Time Scale of Artificial Intelligence
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.3790

Intelligence Explosion

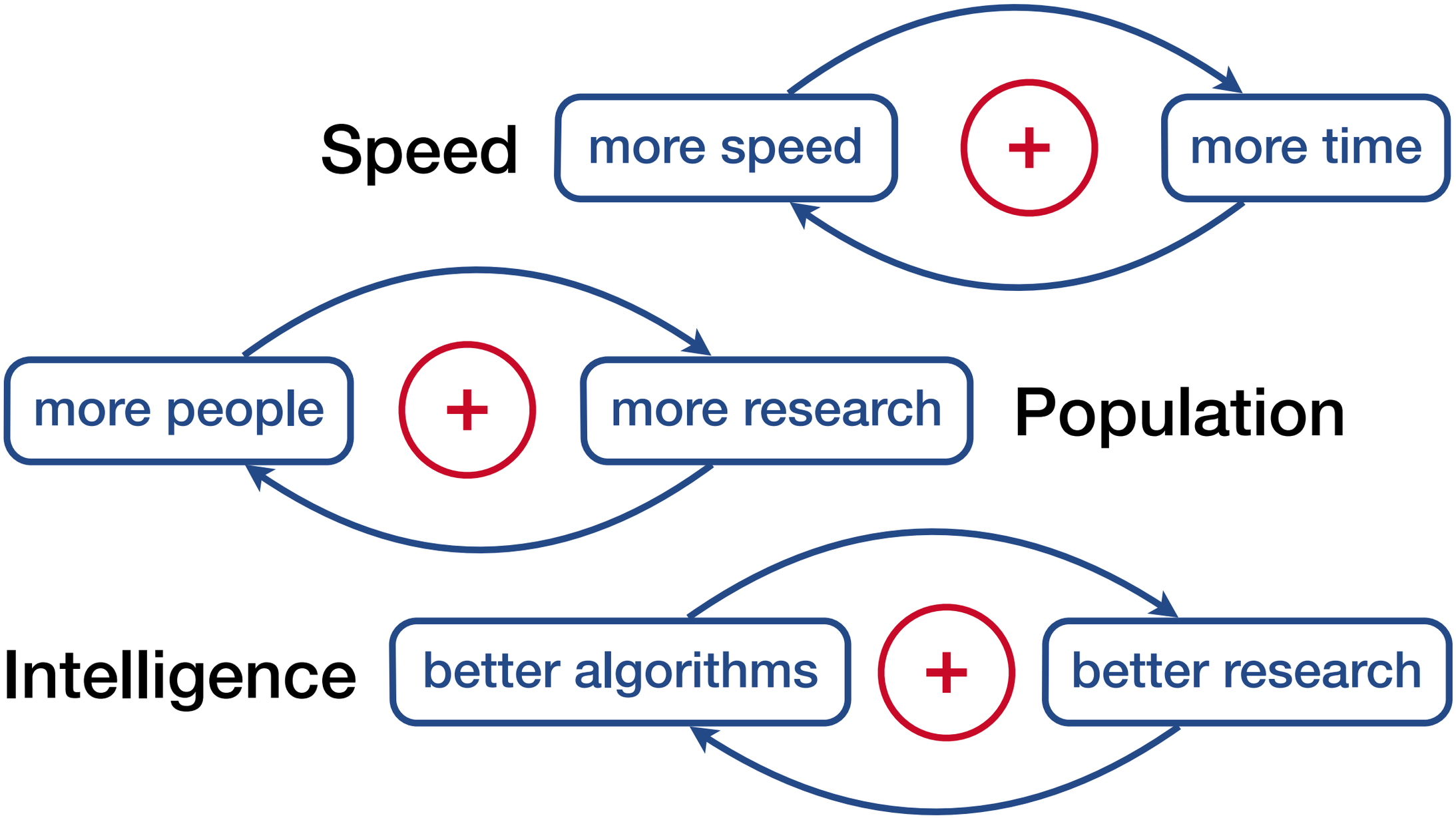
Qualitative

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

Three Separate Explosions



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

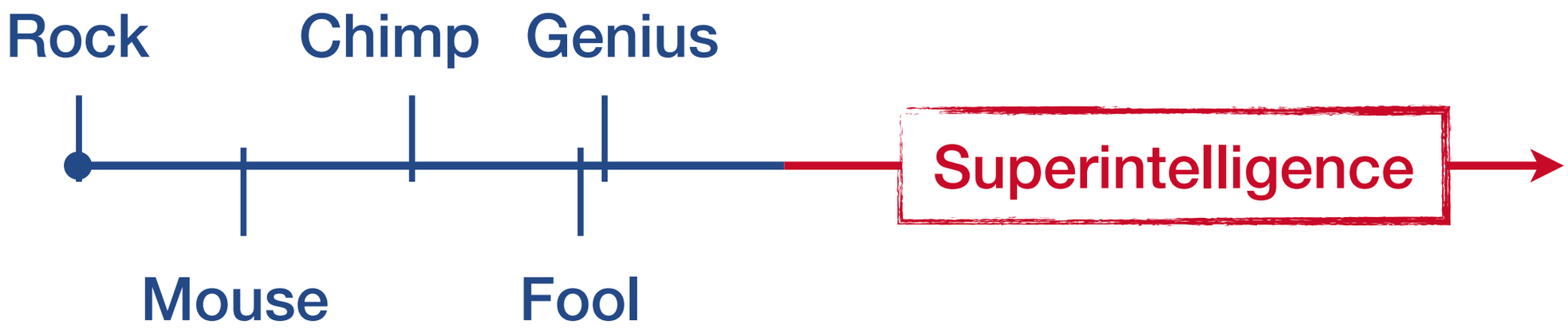
Three Major Singularity Schools
yudkowsky.net/singularity/schools/

Superintelligence

What are potential outcomes?

Definition of Superintelligence

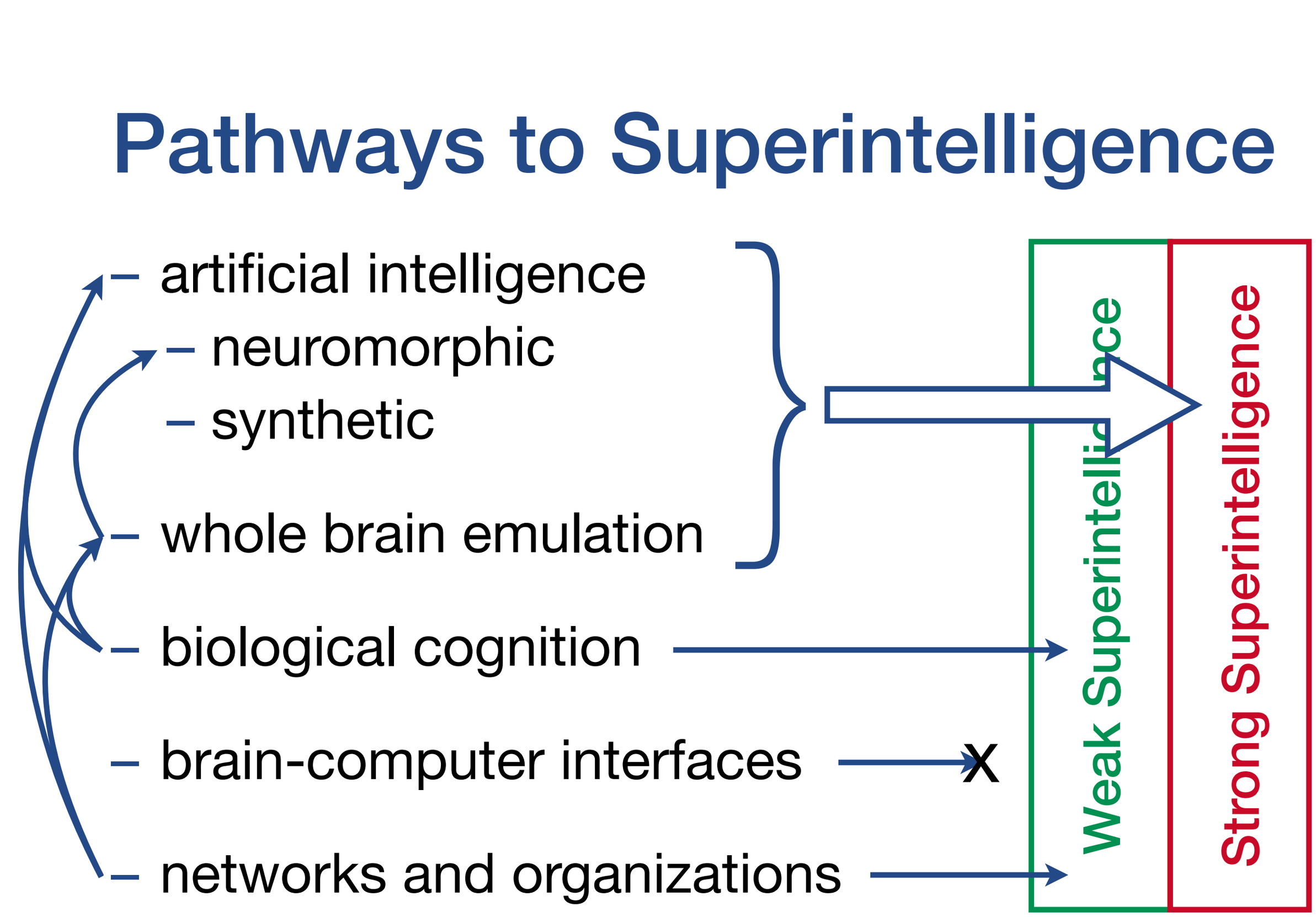
An agent is called **superintelligent** if it exceeds the level of current human intelligence in all areas of interest.



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

Pathways to Superintelligence

- artificial intelligence
  - neuromorphic
  - synthetic
- whole brain emulation
- biological cognition
- brain-computer interfaces
- networks and organizations



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

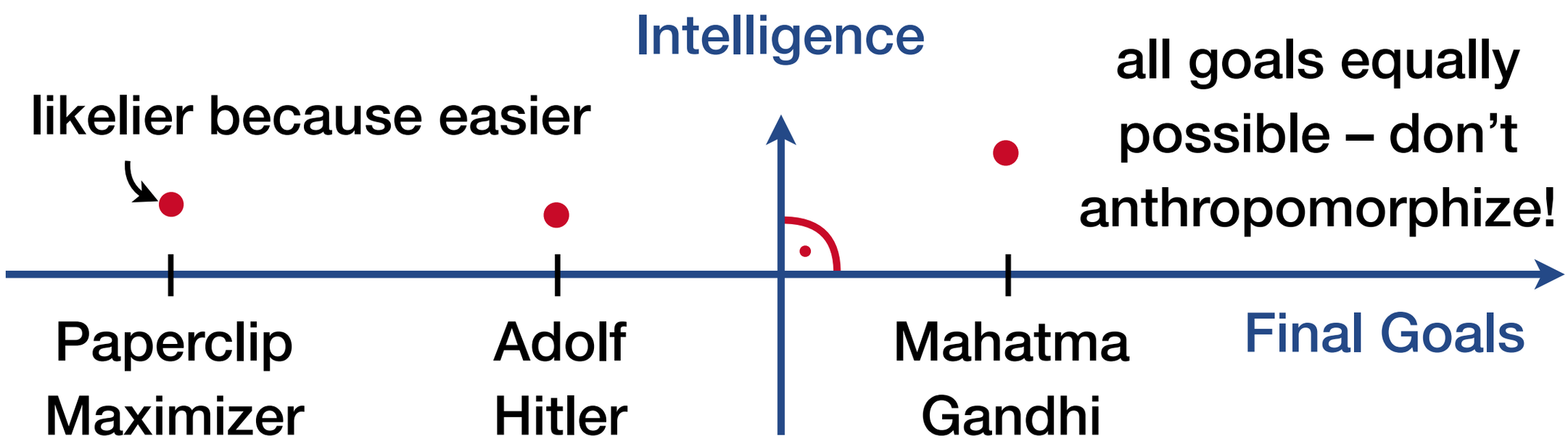
Advantages of AIs, Uploads and Digital Minds
kajsotala.fi/Papers/DigitalAdvantages.pdf

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

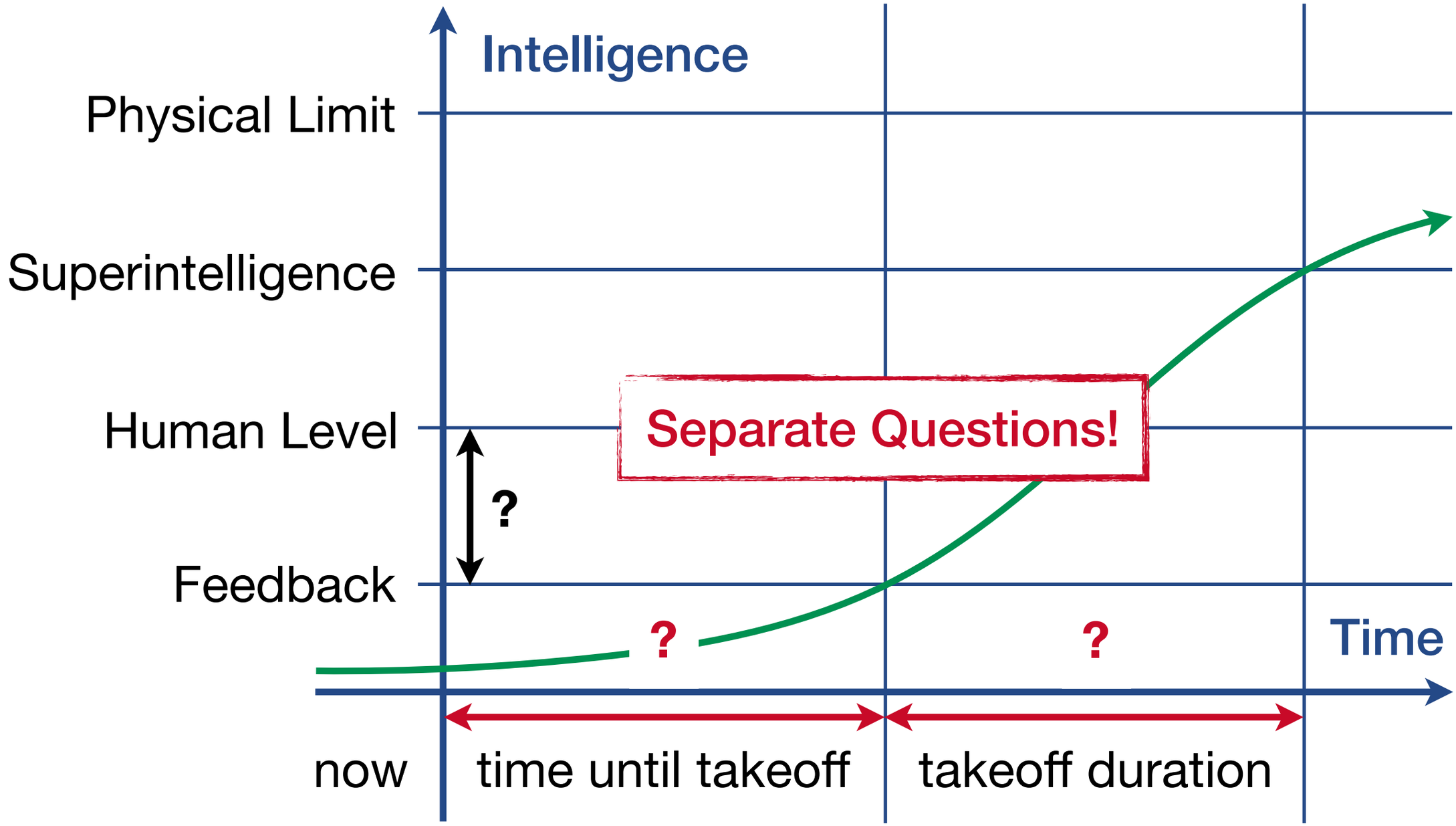
Single-Shot Situation

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 (Slide 9)	**Second Transition** 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
hard	Logic and Thinking	easy

GPU

… but there is massive progress!

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-[…]

Consumer Products

Siri

INTRODUCING
amazon echo

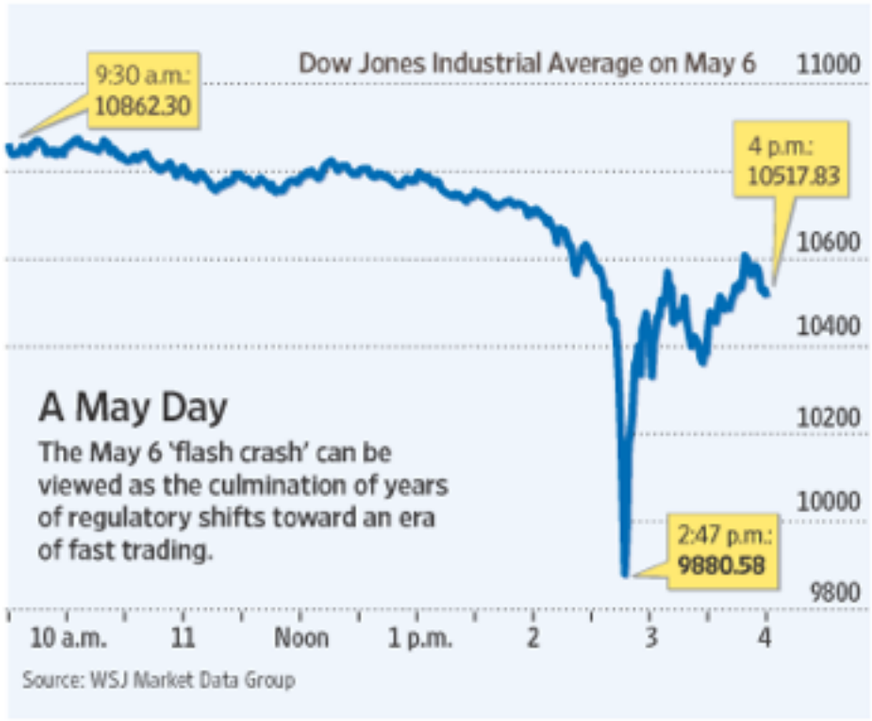
jibo

Military Robots

P.W. Singer
go.ted.com/xe3

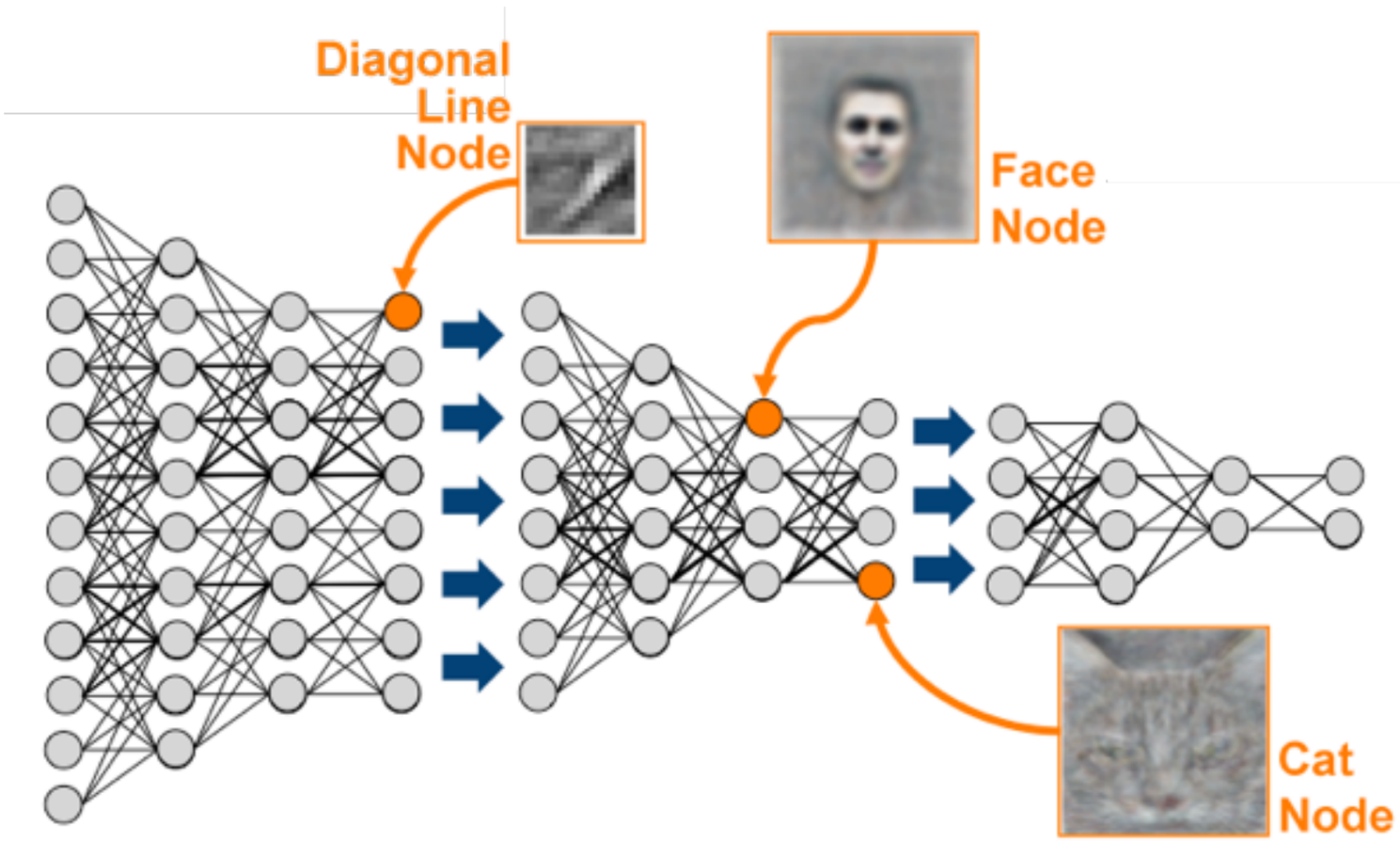
Financial Markets

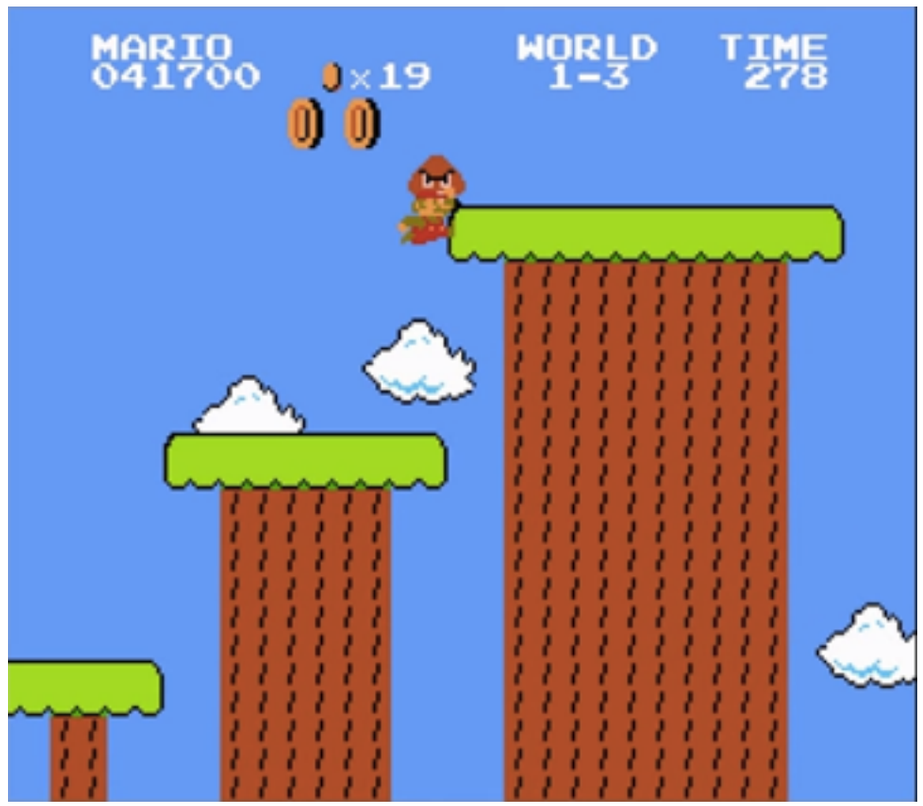
- **High-frequency trading** (HFT): Buy and sell securities within millseconds algorithmically
- In 2009, 65% of all US equity trading volume
- **Flash crash**: very rapid fall in security prices
- 6 May 2010: Dow Jones lost **$1 trillion** (over 9%)
- 23 April 2013: One tweet causes **$136 billion** loss



Kevin Slavin
go.ted.com/xee

Machine Learning





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

Universal Artificial Intelligence

$$a_k := \arg\max_{a_k} \sum_{o_k r_k} ... \max_{a_m} \sum_{o_m r_m} (r_k + ... + r_m) \sum_{q\,:\,U(q,\,a_1..a_m) = o_1 r_1 .. o_m r_m} 2^{-l(q)}$$

- AIXI by Marcus Hutter at IDSIA in Manno
- AIXI is a universally optimal rational agent
- AIXI uses Solomonoff induction and EUT
- AIXI is gold standard but not computable

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

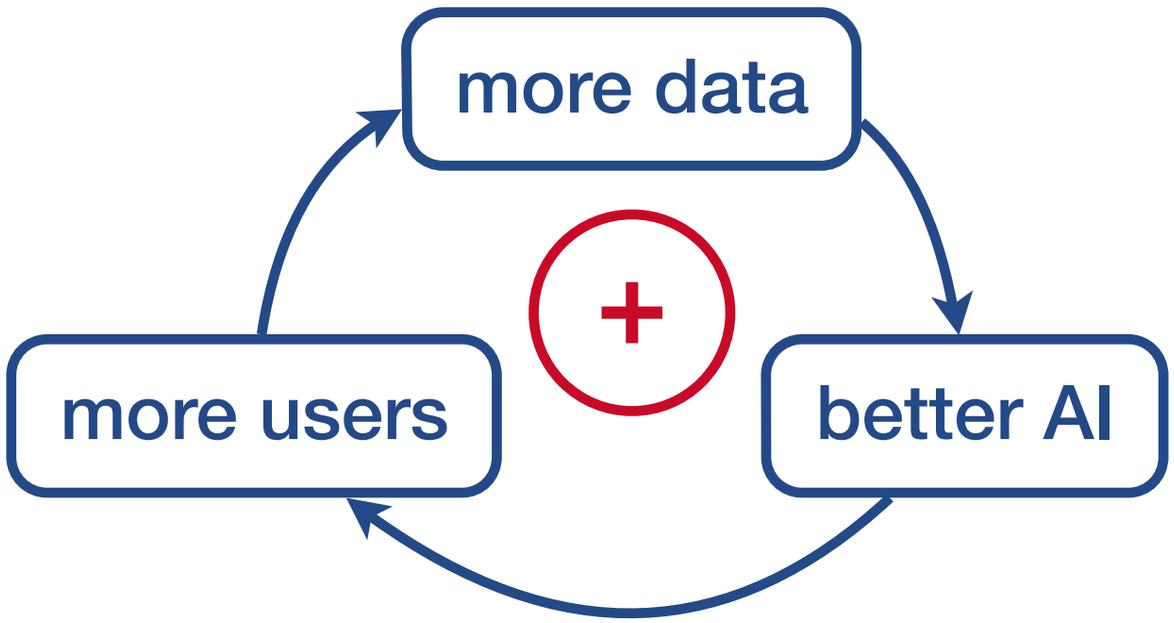
Accelerators

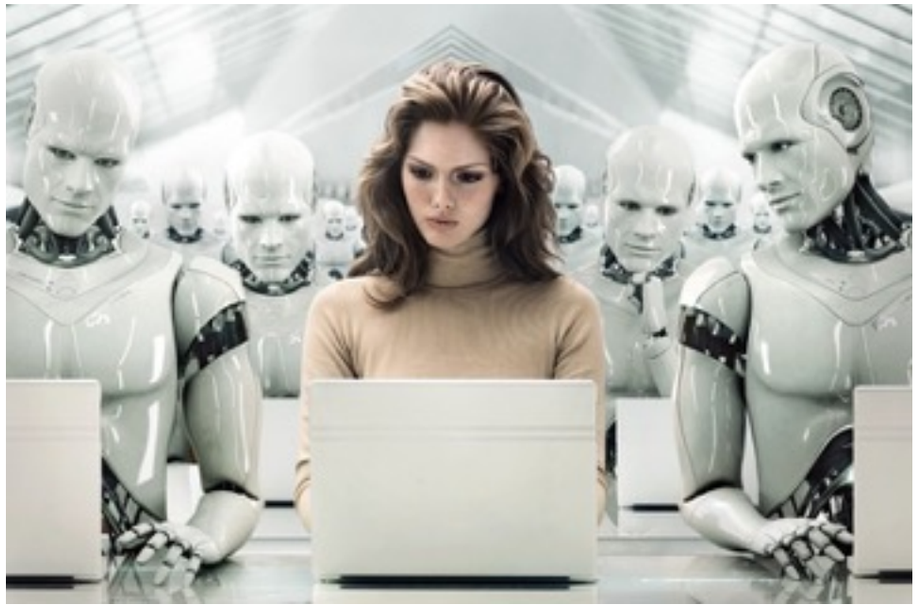
- Faster **hardware**
- Better **algorithms**
- Massive **datasets**

+ enormous incentives!

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

Economic Incentives





- It’s difficult to enter the race later on
- Machines do more intellectual tasks
- Impossible for humans to **compete**

3 Breakthroughs That Have Unleashed AI on the World
www.wired.com/2014/10/future-of-artificial-intelligence/

Economic Consequences

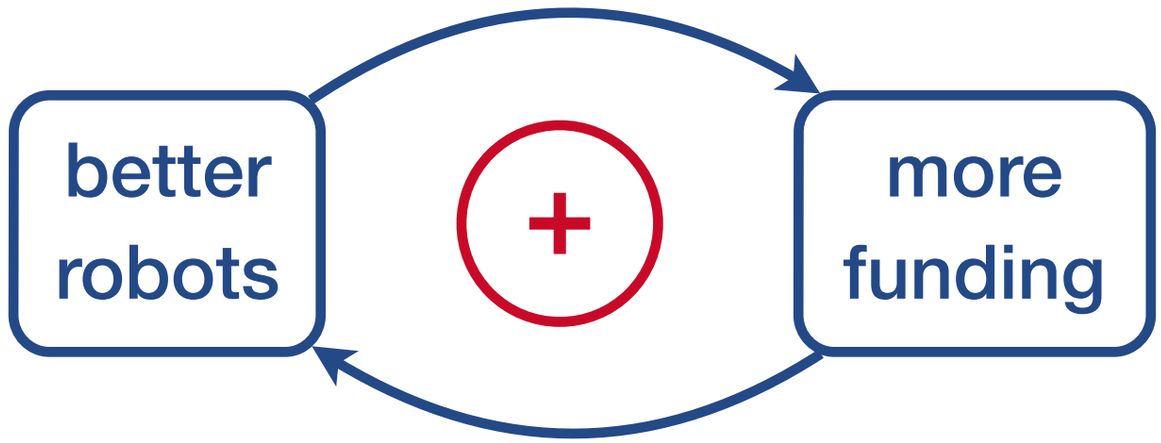
- The living costs of digital workers are **drastically lower** (just energy)
- Thus enormous pressure on wages
- Massive **unemployment** ahead of us
- Wages approach zero, wealth infinity

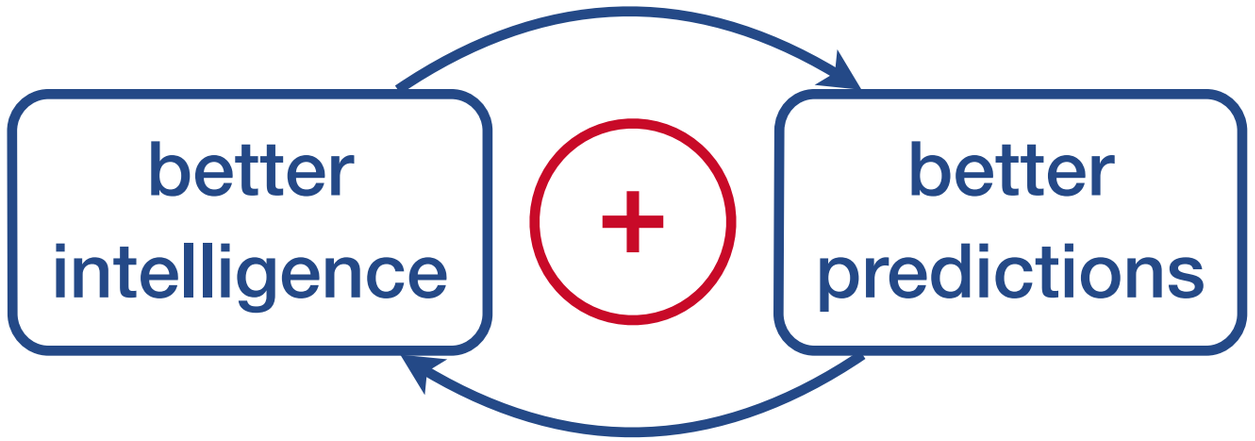
Introduce unconditional basic income?

Humans Need Not Apply
youtu.be/7Pq-S557XQU

Military Incentives

Arms Race?





Egoistic Incentives

- Intelligence
- Wellbeing
- Longevity

⇒ **Willing to take risks**

But with great power comes great responsibility!

PostHuman: An Introduction to Transhumanism
www.youtube.com/watch?v=bTMS9y8OVuY

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

Escaping the Box

The AI could **persuade** someone to free it from its box and thus **human control** by:

- Offering **wealth and power** to liberator
- Claiming it needs **outside resources** to accomplish a task (like curing diseases)
- Predicting a real-world disaster which occurs and claiming afterwards it could have been **prevented** had it been let out

Yudkowsky: The AI-Box Experiment
yudkowsky.net/singularity/aibox/

Value Loading

Utility function of AI?

- Perverse instantiation
- Moral blind-spots…?

Coherent Extrapolated Volition (CEV):

The AI should do what we would want, if we were more intelligent, better informed and more the people we wished we were.

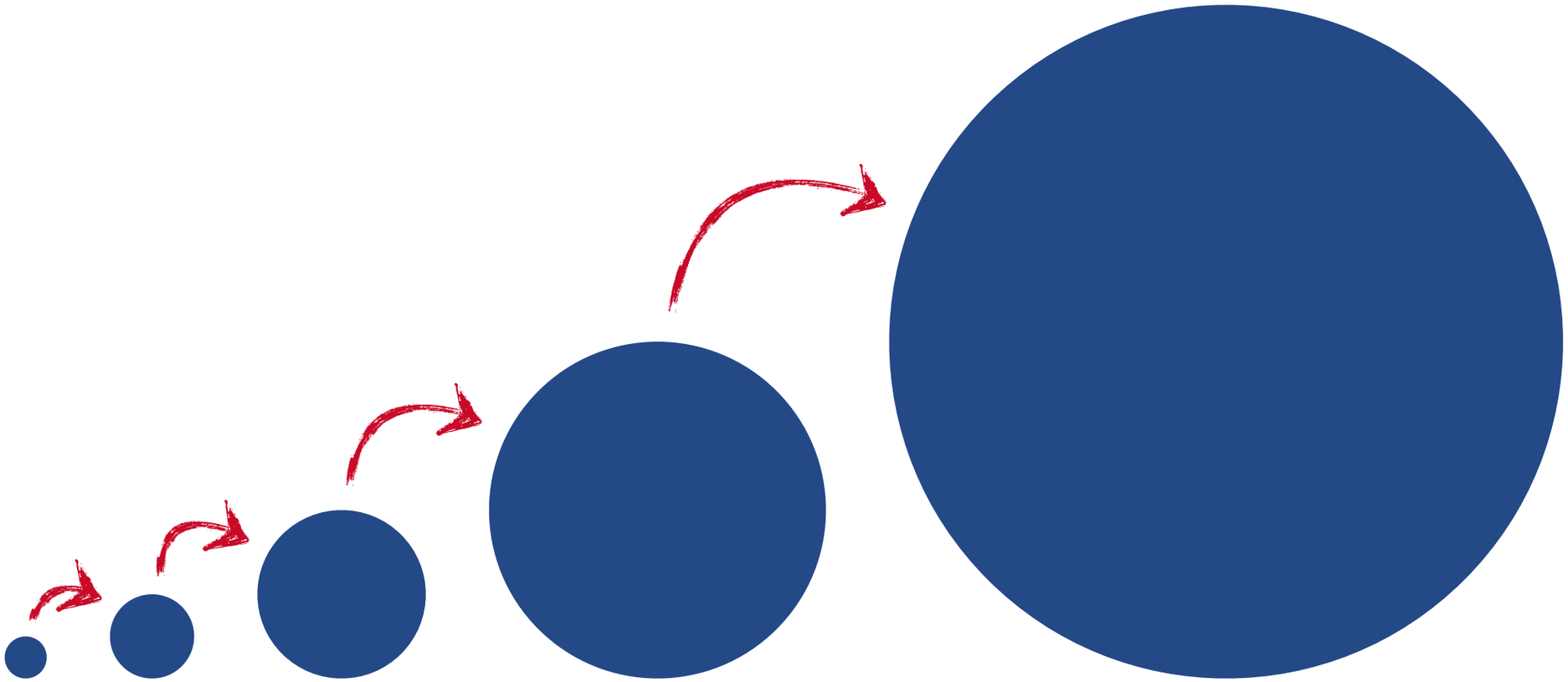
Coherent Extrapolated Volition
intelligence.org/files/CEV.pdf

Goal-Directedness and Tool AI

- **Orthogonality Thesis** (revisited): Any utility function can be combined with a powerful epistemology and decision theory.
- Why not create an AI **without motivations**?
- Boxed **oracle AI** could work but less useful
- AI is relevant to find solutions for problems:
  - **might be unintended** (perverse instant.)
  - **might require planning** to meet criterion

Controlling and Using an Oracle AI
www.nickbostrom.com/papers/oracle.pdf

Stable Self-Improvement



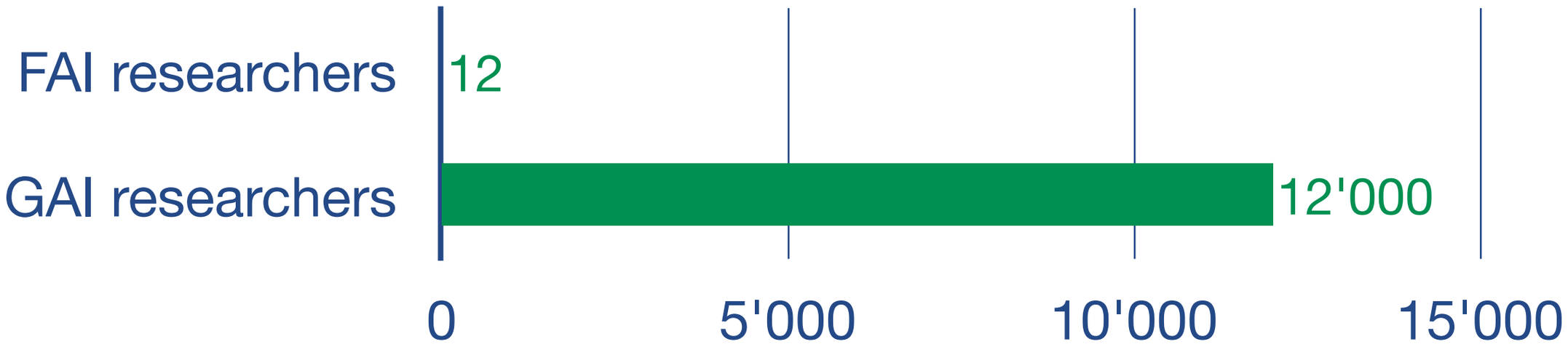
Friendly

Friendly?

Differential Intellectual Progress

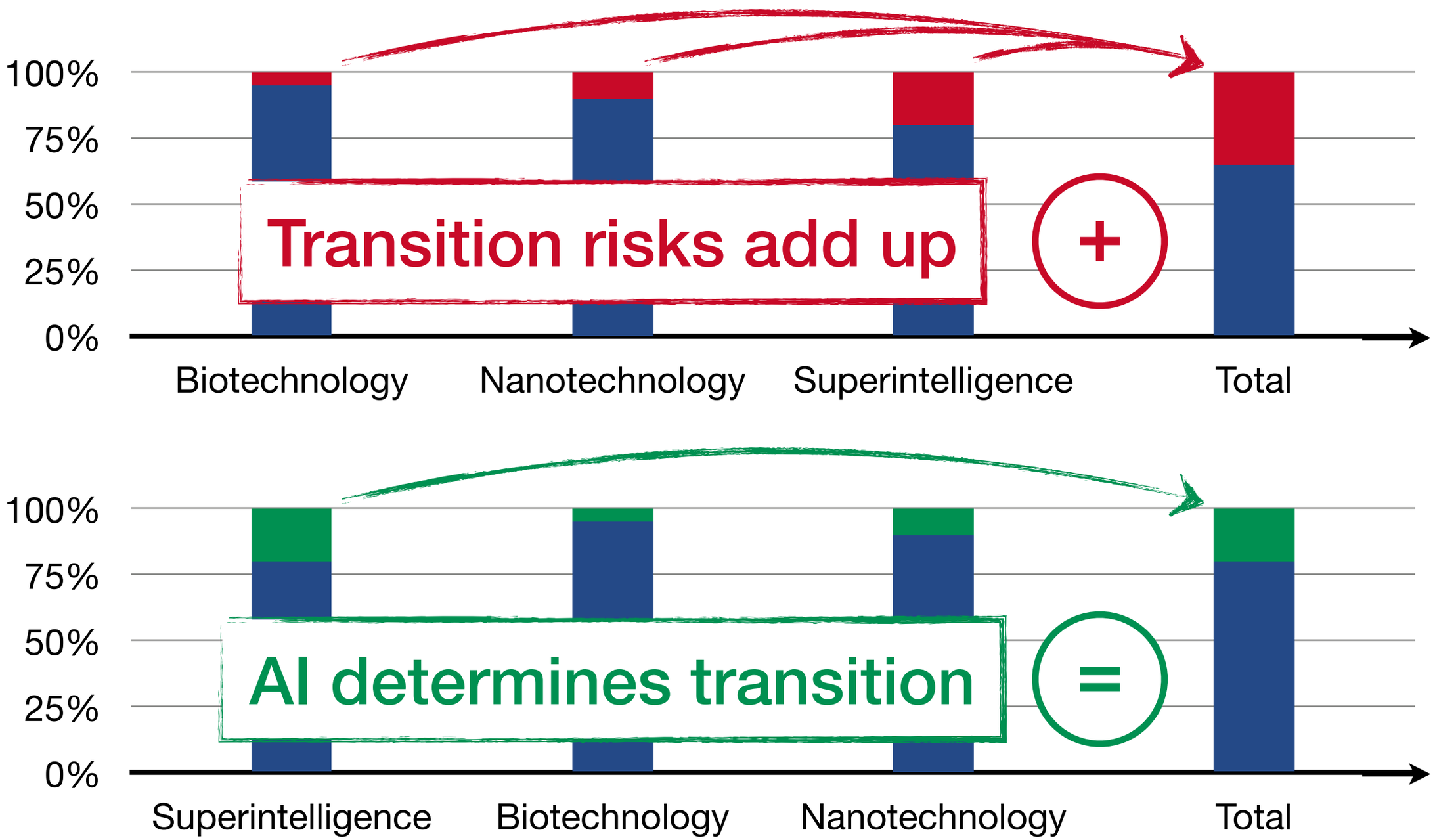
Prioritize risk-reducing intellectual progress over risk-increasing intellectual progress

AI safety should outpace **AI capability** research



Differential Intellectual Progress as a Positive-Sum Project
foundational-research.org/[…]/differential-progress-[…]/

Order of Arrival



Information Hazards

Research can …

– reduce the great uncertainties

… but can also …

– bring up dangerous insights or ideas

Information Hazards: A Typology
www.nickbostrom.com/information-hazards.pdf

Creating Awareness

Outreach can …

– create awareness

… but can also …

– fuel existing fears and cause panic!

80’000 Hours: Professional Influencing
80000hours.org/[…]professional-influencing/

Prisoner's Dilemma

- Difficult to prevent **arms races**
- Parties are better off by **defecting**
- The winner takes all (of what remains)

- Arms races are **dangerous** because

parties sacrifice safety for speed!

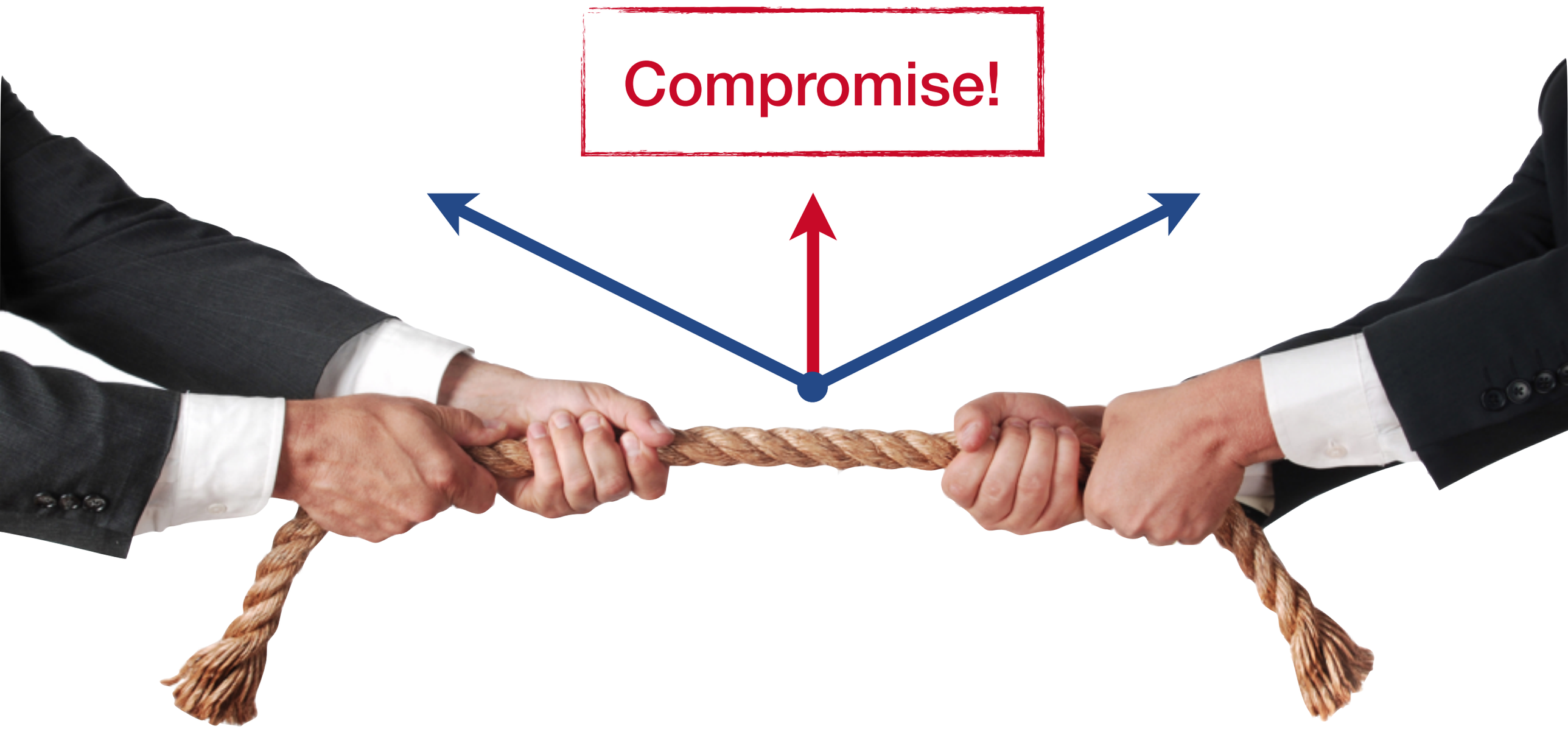
Armstrong, Bostrom, Shulman: Racing to the Precipice
www.fhi.ox.ac.uk/[…]/Racing-to-the-precipice-[…].pdf

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!

Moral Trade



Brian Tomasik: Gains from Trade through Compromise
foundational-research.org/[…]/gains-from-trade-[…]/

Heuristics for Altruists

Safe bets that likely turn out positive:

- **Remain alive!** (Self-Preservation)
- **Remain an altruist!** (Goal-Preservation)
- **Acquire wealth and influence.** (Resource Ac.)
- **Educate yourself and become more rational.** (Self-Improvement, Intelligence Accumulation)

80’000 Hours: Career Guide
80000hours.org/career-guide/

Sources

Where to learn more?

Institutes and Influential People

Future of Humanity Institute
UNIVERSITY OF OXFORD

MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Foundational Research
INSTITUTE

Nick Bostrom

Eliezer Yudkowsky

Brian Tomasik

Talks

The long-term future of AI(and what we can do about it): Daniel Dewey at TEDxVienna

When creative machines overtake man: Jürgen Schmidhuber at TEDxLausanne

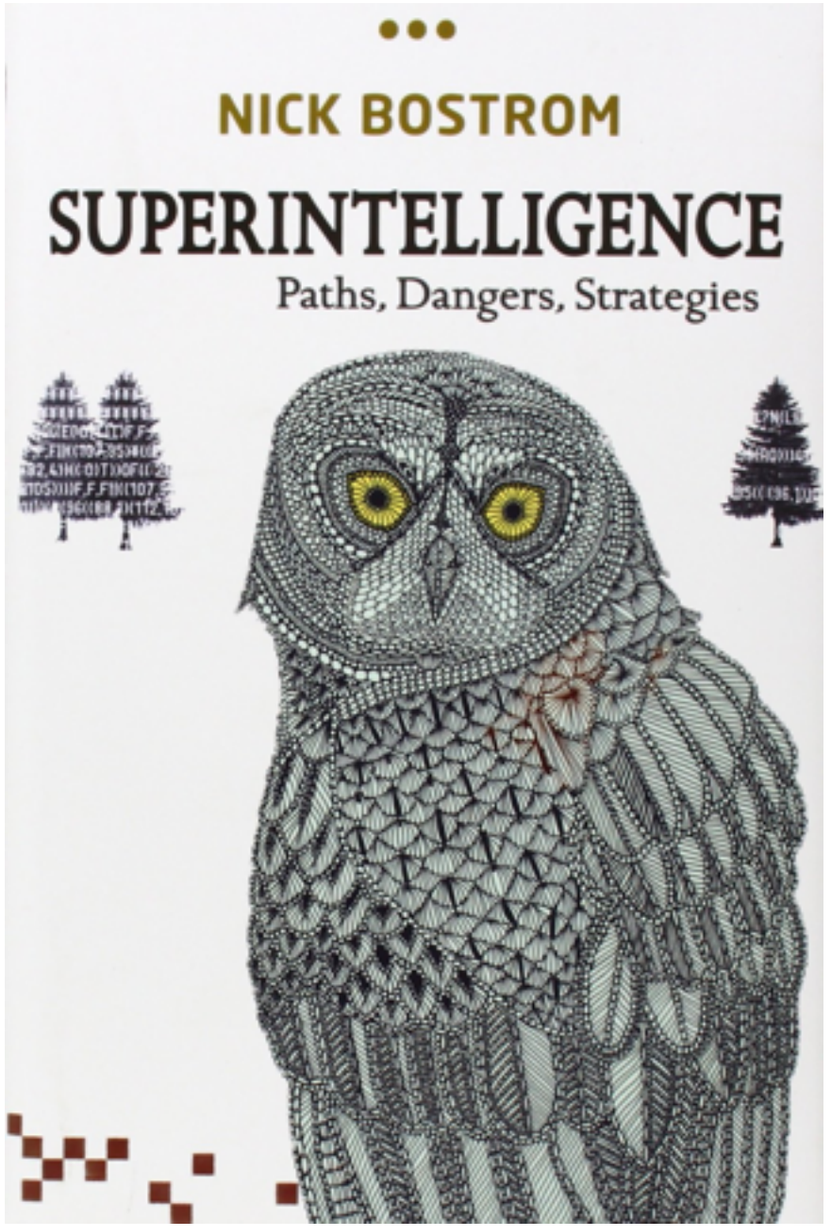
Daniel Dewey
TEDxVienna

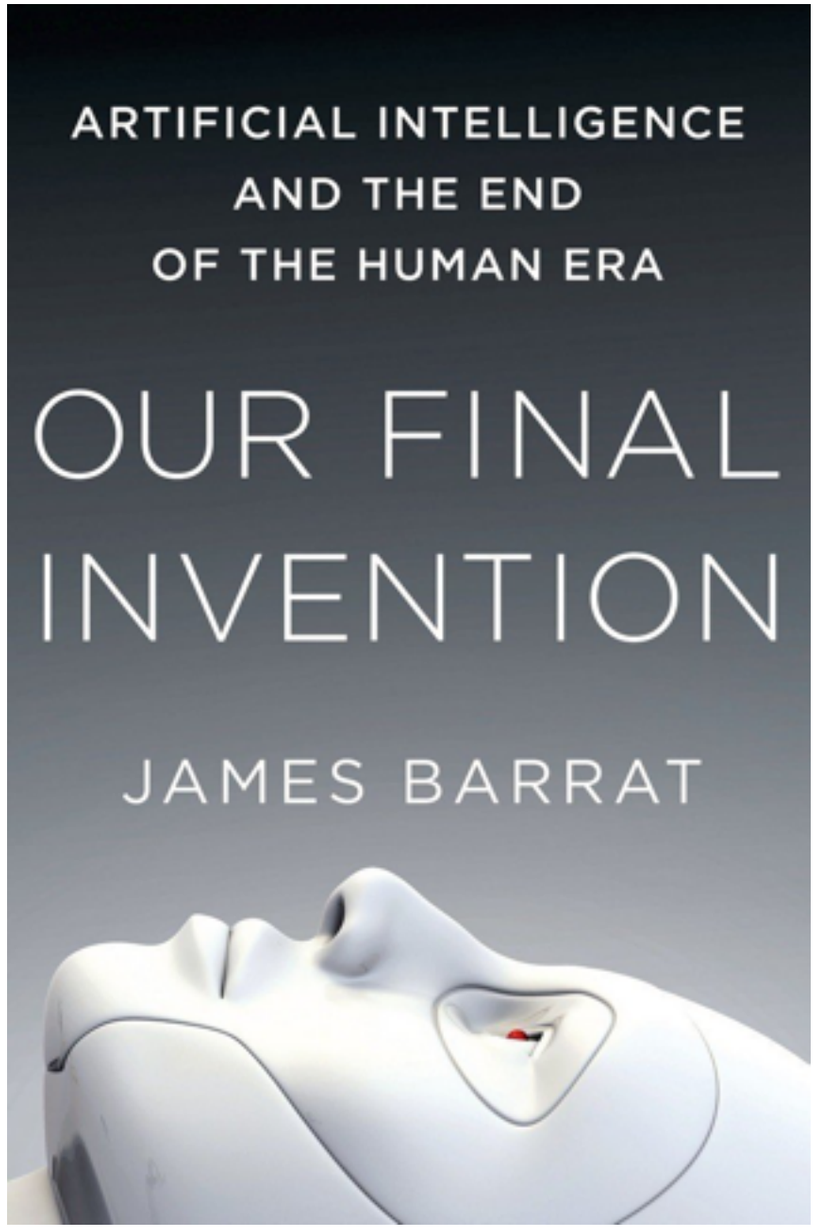
Jürgen Schmidhuber
TEDxLausanne

Papers

- Intelligence Explosion by **Luke Muehlhauser** and **Anna Salamon**
- The Singularity: A Philosophical Analysis by **David Chalmers**
- The Superintelligent Will by **Nick Bostrom**

Books


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies


ARTIFICIAL INTELLIGENCE
AND THE END
OF THE HUMAN ERA
OUR FINAL
INVENTION
JAMES BARRAT

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*

giordano bruno stiftung **schweiz**

Discussion

www.superintelligence.ch

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