Can Intelligence Explode?

Marcus Hutter
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

Australian National University

Abstract

The technological singularity refers to a hypothetical scenario in which technological advances virtually explode. The most popular scenario is the creation of super-intelligent algorithms that recursively create ever higher intelligences. After a short introduction to this intriguing potential future, I will elaborate on what it could mean for intelligence to explode. In this course, I will (have to) provide a more careful treatment of what intelligence actually is, separate speed from intelligence explosion, compare what super-intelligent participants and classical human observers might experience and do, discuss immediate implications for the diversity and value of life, consider possible bounds on intelligence, and contemplate intelligences right at the singularity.

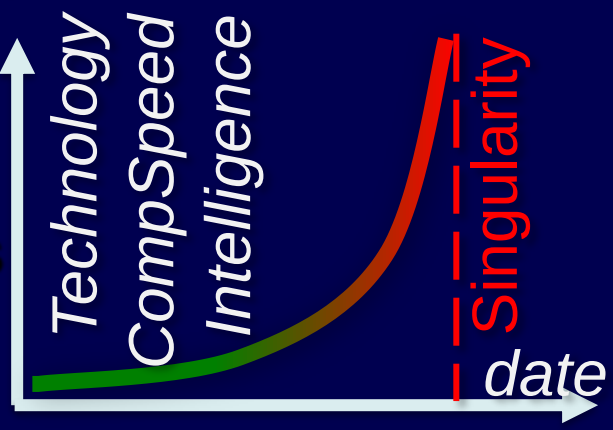
Table of Contents


Introduction to the
Technological Singularity

What is the Technological Singularity

Definition: The Singularity…

- is a hypothetical scenario in which self-accelerating technological advances cause infinite progress in finite time.



Intelligence &| Speed explosion

- (Good 1965; Yudkowsky 1996; Chalmers 2010; …).

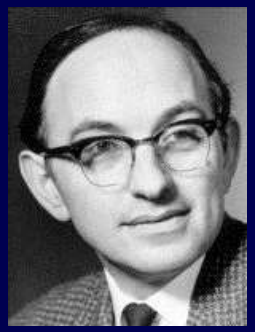
Prediction barrier

- Radically changing society ultimately becomes incomprehensible to us current humans. Still some general aspects may be predictable.

History

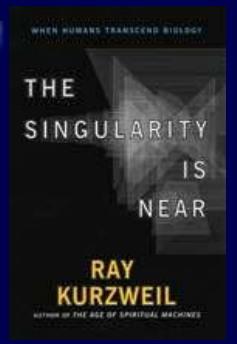
- Ancient (Thornton 1847)
- In science fiction / mathematicians

  Stanislaw Ulam (1958)
  I.J. Good (1965)
  Ray Solomonoff (1985)
  Vernor Vinge (1993)

- Wide-spread popularization

  Kurzweil Books (1999,2005) . Internet.



- Events (Singularity Summit 2006+)
- Organizations (Singularity Institute 2000+ & University)
- Philosophers (David Chalmers 2010)

Related Developments

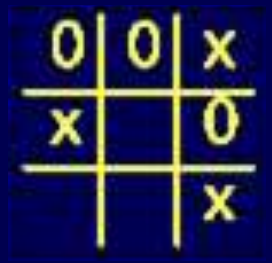
- Artificial General Intelligence
  AGI conference series 2008+
- Whole-brain emulation 10^9€ and $3\times10^9$$ projects
- Universal AI theory of most intelligent agent
- Immortalism extend human life-span ideally indefinitely
- Transhumanism enhancing humans, H+
- Omega Point Universe evolves towards maximum level of complexity and consciousness

Paths to Singularity

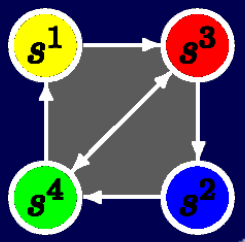
I only consider arguably most plausible scenario of software intelligence based on increasingly powerful hardware.

Still this leaves many options, the major ones being:

- mind uploading (via brain scan) & subsequent improvement
- knowledge-based reasoning and planning software (traditional AI research)



- artificial agents that learn from experience (the machine learning approach)



- self-evolving intelligent systems (genetic algorithms and artificial life approach)

- awakening of the Internet (digital Gaia scenario).

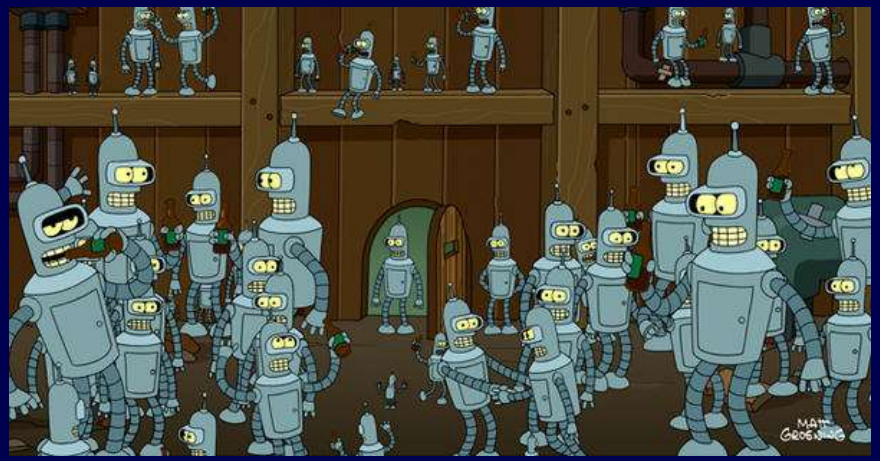
- ~~brain enhancement technologies (drugs genetic engineering)~~

Considered Setup

- virtual software society consisting of interacting rational agents whose intelligence is high enough to construct the next generation of more intelligent rational agents.

- I will discuss what (super)intelligence and rationality could mean in this setup.

- For concreteness, envisage an initial virtual world similar to our current real world and inhabited by human mind uploads.



Terminology & Jargon

- **comp** = computational resources
- **singularity** = infinite change of an observable quantity in finite time
- **intelligence explosion** = rapidly increasing intelligence far beyond human level
- **intelligence singularity** = infinite intelligence in finite time
- **speed explosion/singularity** = rapid/infinite increase of computational resources
- **outsider** = biological = non-accelerated real human watching a singularity
- **insider** = virtual = software intelligence participating in a singularity
- **computronium** = theoretically best possible computer per unit of matter
- **real/true intelligence** = what we intuitively would regard as intelligence
- **numerical intelligence** = numerical measure of intelligence like IQ score
- **AI** = artificial intelligence (used generically in different ways)
- **AGI** = artificial general intelligence = general human-level intelligence or beyond.
- **super-intelligence** = AI+ = super-human intelligence
- **hyper-intelligent** = AI++ = incomprehensibly more intelligent than humans
- **vorld** = virtual world. A popular oxymoron is `virtual reality'
- **virtual** = software simulation in a computer

Global (Simplifying) Assumption

Strong Church-Turing Thesis

All physical processes, including the human mind and body, are computational and can be simulated (virtualized) by a sufficiently powerful (theoretical) computer.

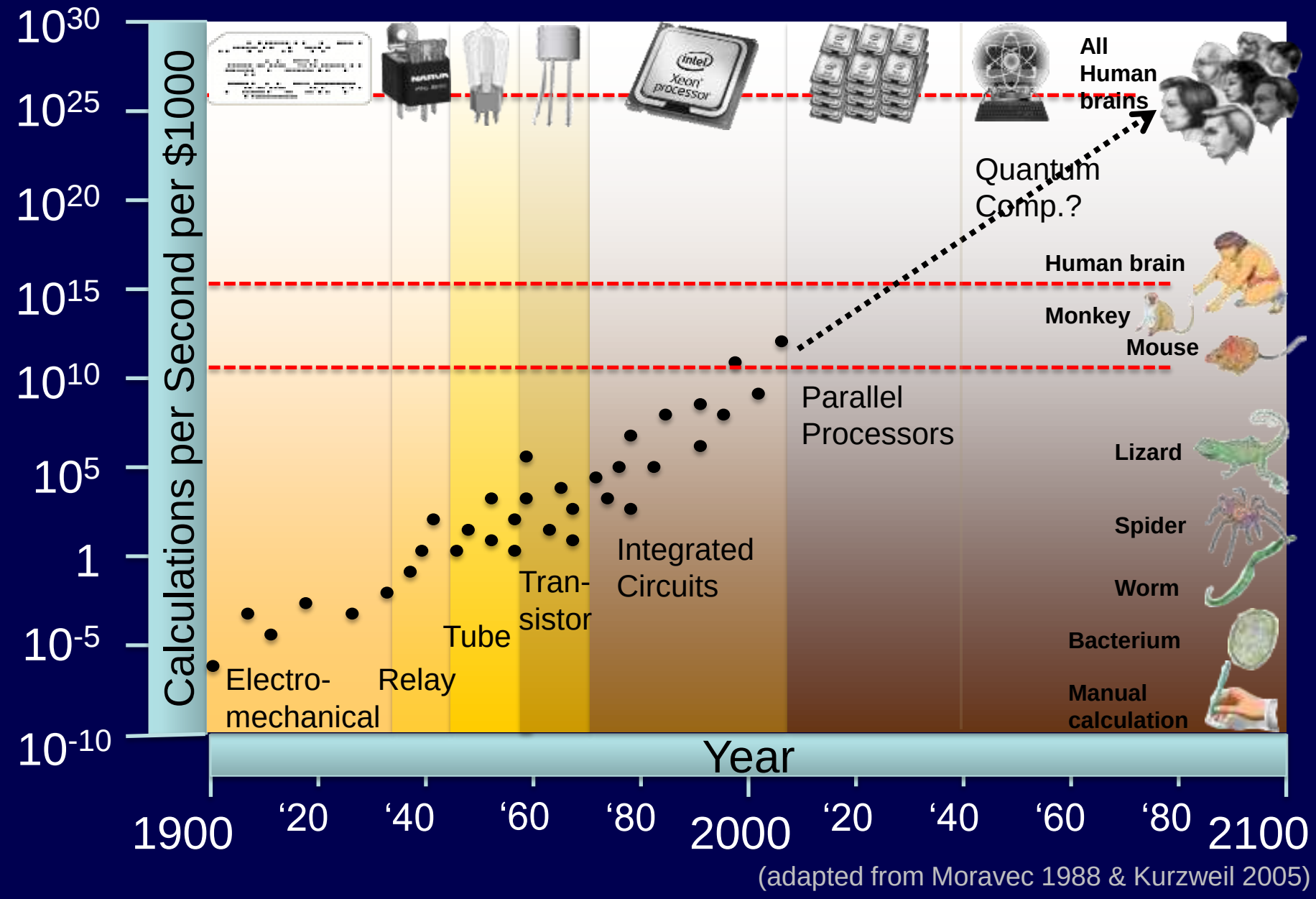
Justifications:

- Deutsch (1997)
- Rathmanner & Hutter (2011)
- Chalmers (2010)

… and many others …


Will there be
a Singularity
?

Moore's Law



(adapted from Moravec 1988 & Kurzweil 2005)

Super-Intelligence by Moore's Law

- Moore's law: comp doubles every 1.5yrs. Now valid for >50yrs
- As long as there is demand for more comp,
  Moore's law could continue to hold
  for many more decades before computronium is reached.

$\Rightarrow$ in 20-30 years the raw computing power of a single computer will reach 10^{15}...10^{16} flop/s.

- Computational capacity of a human brain: 10^{15}...10^{16} flop/s
- Some Conjecture: software will not lag far behind
  (AGI *or* reverse engineer *or* simulate human brain)

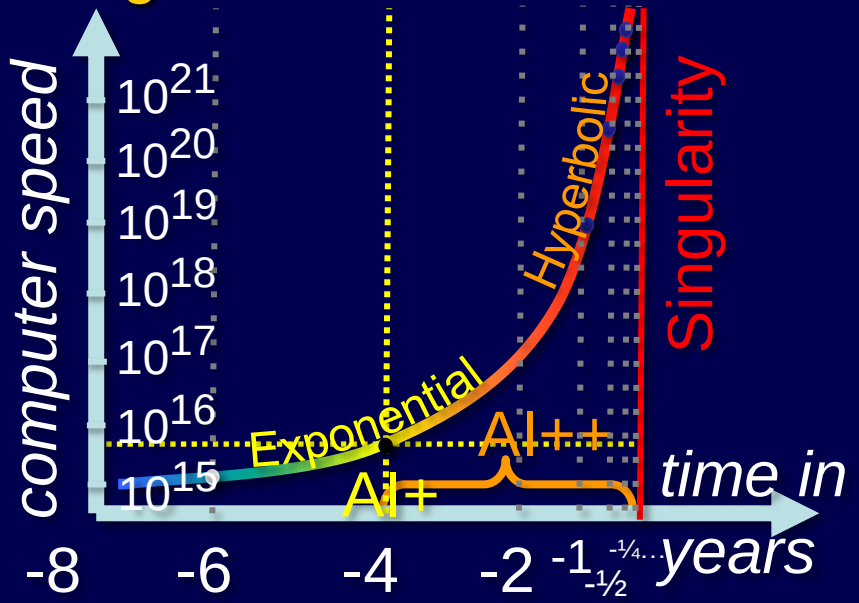
$\Rightarrow$ **human-level AI in 20-30 years?**

Singularity by Solomonoff's Law

If computing speeds double every two years, what happens when computer-based AIs are doing the research?

- Computing speed doubles every two years.
- Computing speed doubles every two years of work.
- Computing speed doubles every two subjective yrs of work.
- Two years after Artificial Intelligences reach human equivalence, their speed doubles.
- One year later, their speed doubles again.
- Six months - three months - 1.5 months ... Singularity.

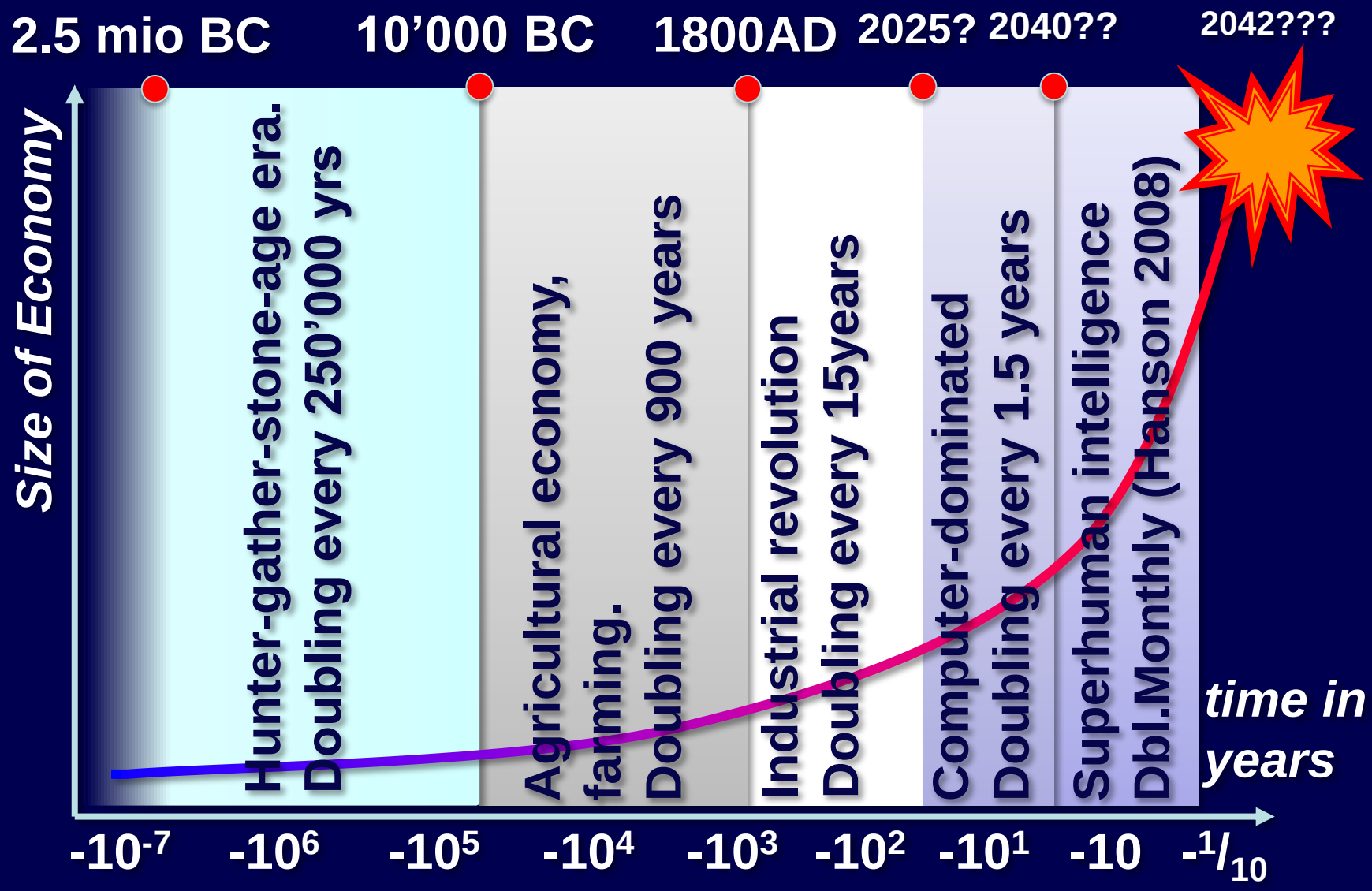
(Yudkowski 1996)



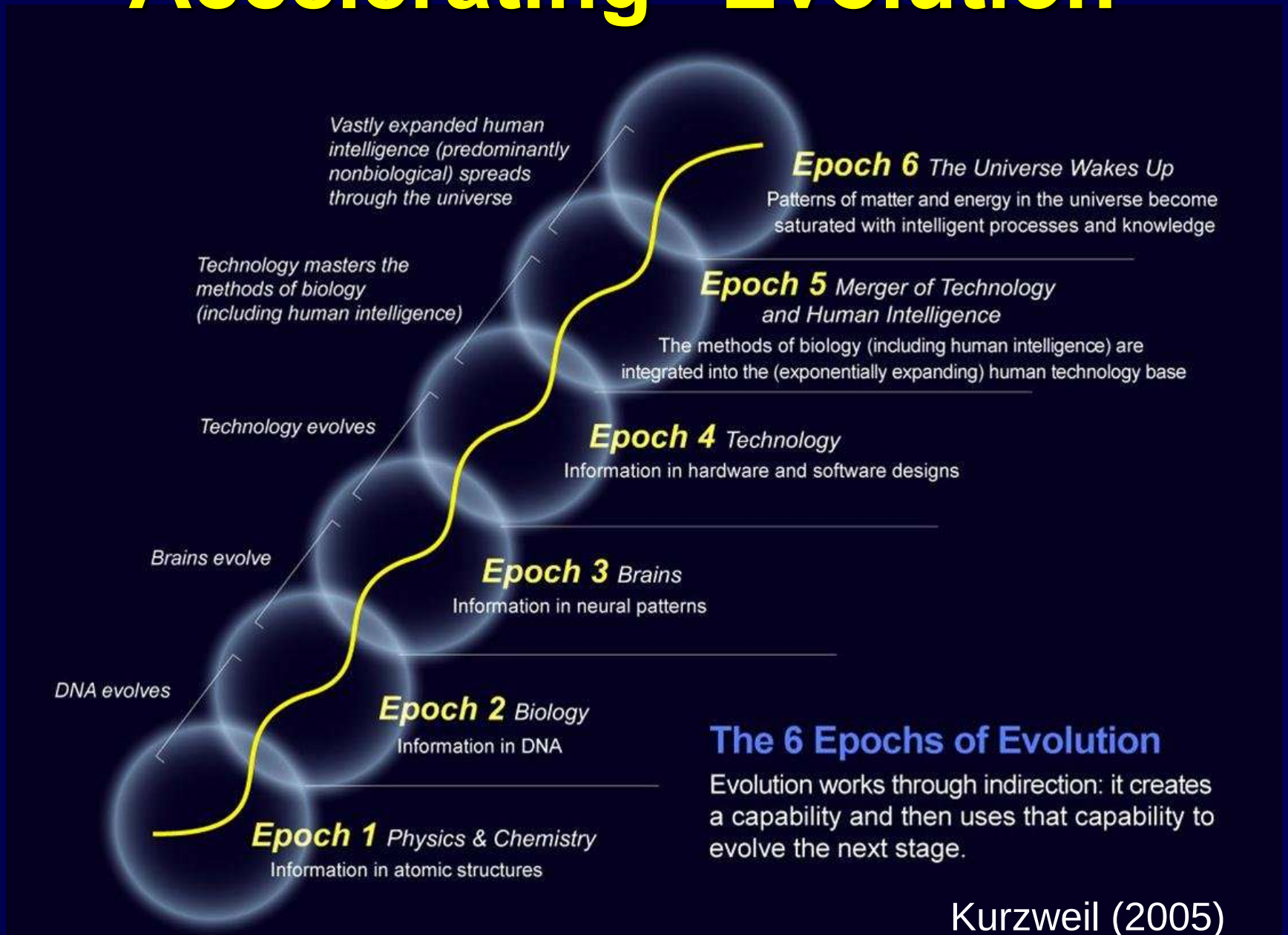
Moore's law predicts its own break-down!

But not the usually anticipated slow-down, but an enormous acceleration of progress when measured in physical time.

Acceleration of Doubling Patterns



Accelerating "Evolution"



Kurzweil (2005)

Obstacles Towards a Singularity

- Structural obstacles: limits in intelligence space, failure to takeoff, diminishing returns, local maxima
- Manifestation obstacles: disasters, disinclination, active prevention
- Correlation obstacles: speed | technology ↔ intelligence, ...
- Physical limits: Bremermann (1965)(quantum) limit: 10^{50} bits/kg/s
  Bekenstein (2003)(black hole) bound: 10^{43} bits/kg/m or 10^{69} bits/m^2
- But: converting our planet into computronium would still result in a vastly different vorld, which could be considered a reasonable approximation to a true singularity.

- Hard&Software Engineering difficulties: many
- But: Still one or more phase transitions a la Hanson may occur.
- Disinclination to create it:
  most (but not too) likely defeater of a singularity (Chalmers 2010)

Is the Singularity Negotiable

- Appearance of AI+ = ignition of the detonation cord towards the Singularity = point of no return
- Maybe Singularity already now unavoidable?
- Politically it is very difficult (but not impossible) to resist technology or market forces

⇒ it would be similarly difficult to prevent AGI research and even more so to prevent the development of faster computers.

- Whether we are before, at, or beyond the point of no return is also philosophically intricate as it depends on how much free will one attributes to people and society.
- Analogy 1: politics & inevitability of global warming
- Analogy 2: a spaceship close to the event horizon might in principle escape a black hole but is doomed in practice due to limited propulsion.

The Singularity from the Outside

Questions

- What will observers who do not participate in the Singularity “see”?
- How will it affect them?

Terminology

- outsider = biological = non-accelerated real human watching a singularity
- insider = virtual = software intelligence participating in a singularity

Converting Matter into Computers

- The hardware (computers) for increasing comp must be manufactured by (real) machines/robots in factories.
- Insiders will provide blueprints to produce better computers&machines that themselves produce better computers&machines ad infinitum at an accelerated pace.
- Non-accelerated real human (outsiders) will play a diminishing role in this process due to their cognitive and speed limitations.
- Quickly they will only be able to passively observe some massive but incomprehensible transformation of matter going on.

Outward Explosion

- an increasing amount of matter is transformed into computers of fixed efficiency.
- Outsiders will soon get into resource competition with the expanding computer world.
- Expansion rate will approach speed of light so that escape becomes impossible, ending or converting the outsiders' existence.

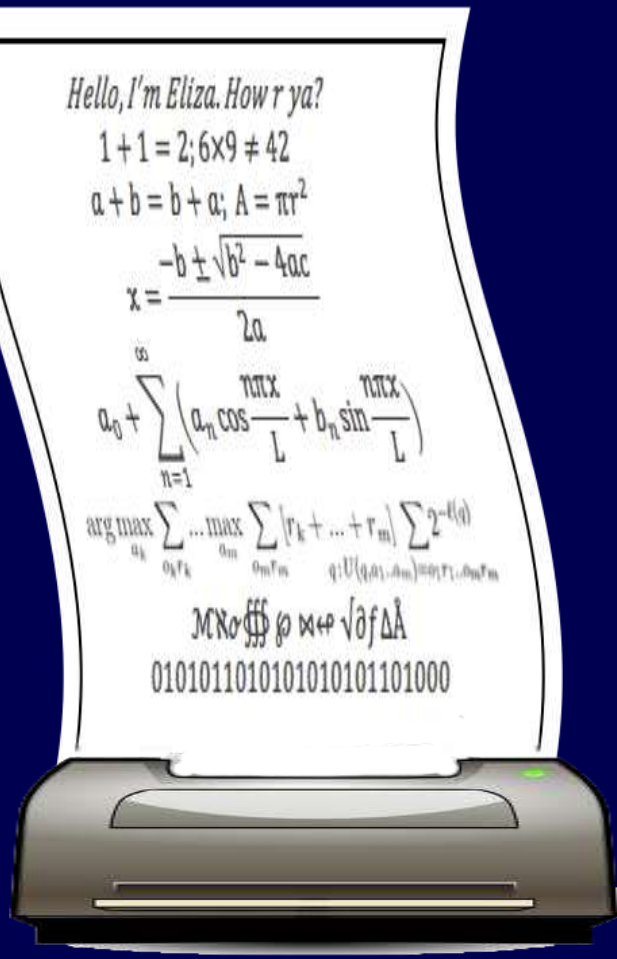
$\Rightarrow$ there will be no outsiders around to observe a singularity

Inward Explosion

- A fixed amount of matter is transformed into increasingly efficient computers.
- Speed of virtual society will make them incomprehensible to the outsiders.
- At best some coarse statistical or thermodynamical properties could ultimately be monitored.

⇒ After a brief period, intelligent interaction between insiders and outsiders becomes impossible.

⇒ outsiders will not experience a singularity



- Even high-speed recording, slowmo communication, or brain augmentation, will not change this conclusion.

Some Information Analogies

- Inside process resembles a radiating black hole observed from the outside.

- Maximally compressed information is indistinguishable from random noise.
- Too much information collapses:
  A library that contains all possible books has zero information content.

Library of Babel: all information = no information

- Maybe a society of increasing intelligence will become increasingly indistinguishable from noise when viewed from the outside.

Comparison

- Each way, outsiders cannot witness a true intelligence singularity.
- Expansion (inward↔outward) usually follows the way of least resistance.
- Inward explosion will stop when computronium is reached.
- Outward explosion will stop when all accessible convertible matter has been used up.
- Historically, mankind was always outward exploring; just in recent times it has become more inward exploring (miniaturization & virtual reality).

The Singularity from the Inside

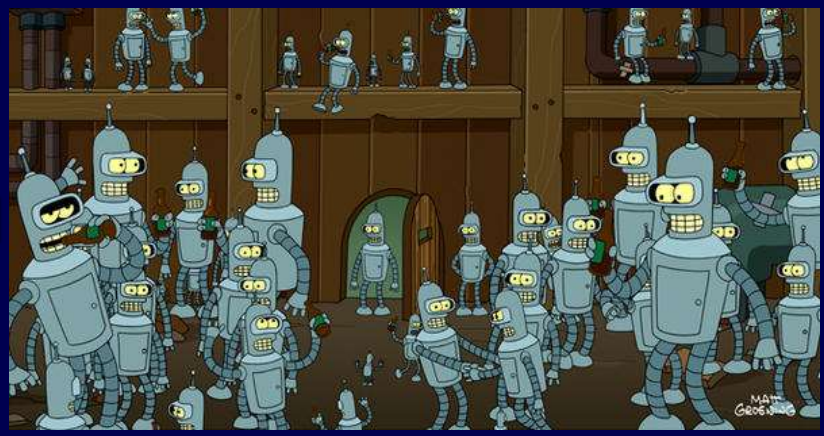
Virtualize Society

Now consider the Singularity from the inside:
What will a participant experience?

Assumptions:

- initial society similar to our current society
- very large number of individuals,
- who possess some autonomy and freedom,
- who interact with each other and with their environment
- in cooperation and in competition over resources and other things.

Example: virtual world populated with intelligent agents simulating scans of human brains.



Fixed Computational Resources

Vorld much like real counter-part: new (virtual) inventions, technologies, fashions, interests, art, etc.

Some difference to real counter-part:

- duplicating (virtual) objects and directed artificial evolution will be easier.
- building faster virtual computers and fancier gadgets will be hard/impossible.
- Virtuals will adapt to abundance/scarcity of virtual resources like in reality
- and/or adapt to new models of society and politics.
- But an intelligence explosion with fixed comp, even with algorithmic improvements seems highly implausible.

Increasing Comp (per Individual)

Assume uniform speed-up of the whole virtual world

- Virtual's subjective thought processes will be sped up at the same rate as their virtual surroundings.
- Then inhabitants would actually not be able to recognize this since nothing would change for them.
- Only difference: outside world slows down.
- Also outsiders would appear slower (but not dumber).
- If comp is sped up hyperbolically, the subjectively infinite future of the virtuals would fit into finite real time.
- Reverse to time dilatation in black holes:
  Astronaut hits singularity in finite/infinite subjective/observer time.

Increasing Comp (# of Individuals)

add more virtuals but keep comp per individual fixed

No individual speedup $\Rightarrow$ intelligence stays bounded

But larger societies can also evolve

- faster (more inventions per real time unit),
- and if regarded as a super-organism,
  there might be an intelligence explosion.

Counter argument: number of individuals involved in a decision process may not be positively correlated with the intelligence of their decision.

Counter examples: Ant colonies and bee hives.

Singularity?

Generalization

Diversity of intelligences

- faster and slower ones,
- higher and lower ones,
- and a hierarchy of super-organisms and sub-vorlds.

Analysis becomes more complicated,
but the fundamental conclusion doesn't change.

Conclusion

Strict intelligence singularity neither experienced by insiders nor by outsiders.

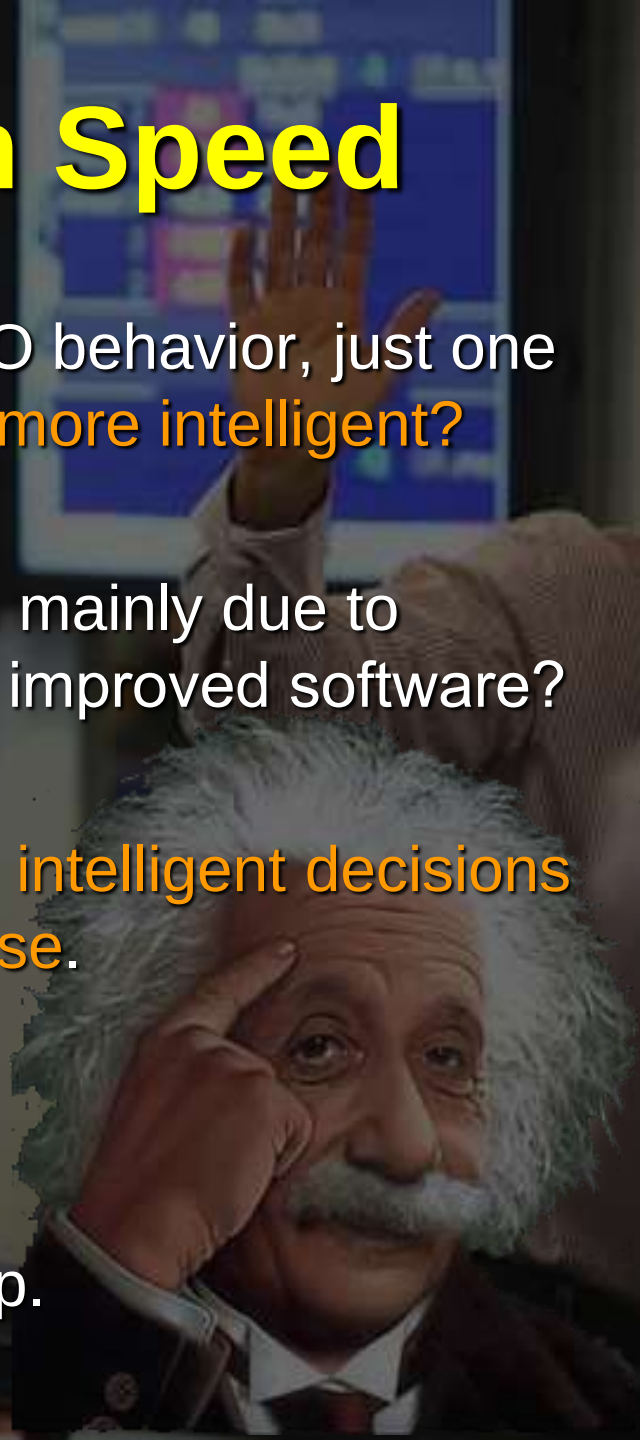
Assume recording technology does not break down:

- then a singularity seems more interesting for outsiders than for insiders.
- On the other hand, insiders actively “live” potential societal changes,
  while outsiders only passively observe them.


Speed Explosion
Intelligence Explosion

Some Thoughts on Speed

- If two agent algorithms have the same I/O behavior, just one is faster than the other, is the faster one more intelligent?
- Has progress in AI …
  improved hardware …
  … been mainly due to
  … or to improved software?
- More comp only leads to more …
  … intelligent decisions
  if the decision algorithm puts it to good use.
- Many algorithms in AI are so-called
  anytime algorithms that indeed
  produce better results if given more comp.



Infinite Comp

- In the limit of infinite comp, in simple and well-defined settings (usually search and planning problems), some algorithms can produce optimal results.
- But for more realistic complex situations (usually learning problems), they saturate and remain sub-optimal.
- But there is one algorithm, namely AIXI, that is able to make optimal decisions in arbitrary situations given infinite comp.
- Fazit:

It is non-trivial to draw a clear boundary between speed and intelligence.

Speedup / Slowdown Effects

Performance per unit real time:

- Speed of agent positively correlates with cognition and intelligence of decisions
- Speed of environment positively correlates with informed decisions

Perf. per subjective unit of agent time from agent's perspective:

- slow down environment = increases cognition and intelligence but decisions become less informed
- speed up environment = more informed but less reasoned decisions

Performance per environment time from env. perspective:

- speed up agent = more intelligent decisions
- slow down agent = less intelligent decisions

Speedup / Slowdown AI Limits

- there is a limit on how much information a comp-limited agent can usefully process or even search through.

- there might also be a limit to how much can be done with and how intelligent one can act upon a limited amount of information.


What
is
Intelligence

What is Intelligence?

- There have been numerous attempts to define intelligence.
- Legg & Hutter (2007) provide a collection of 70+ definitions
  - from the philosophy, psychology, and AI literature
  - by individual researchers as well as collective attempts.
- If/since intelligence is not (just) speed, what is it then?
- What will super-intelligences actually do?

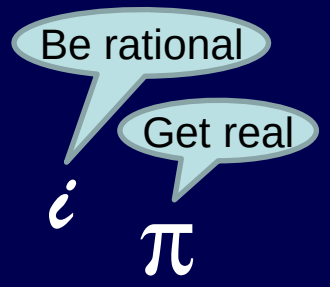
Evolving Intelligence

- Evolution: Mutation, recombination, and selection increases intelligence if useful for survival and procreation.
- Animals: higher intelligence, via some correlated practical cognitive capacity, increases the chance of survival and number of offspring.
- Humans: intelligence is now positively correlated with power and/or economic success (Geary 2007) and actually negatively with number of children (Kanazawa 2007).
- Memetics: Genetic evolution has been largely replaced by memetic evolution (Dawkins 1976),
  the replication, variation, selection, and spreading of ideas causing cultural evolution.

What Activities are Intelligent?
Which Activities does Evolution Select for?

- Self-preservation?
- Self-replication?
- Spreading? Colonizing the universe?
- Creating faster/better/higher intelligences?
- Learning as much as possible?
- Understanding the universe?
- Maximizing power over men and/or organizations?
- Transformation of matter (into computronium?)?
- Maximum self-sufficiency?
- The search for the meaning of life?

Intelligence ≈ Rationality ≈ Reasoning Towards a Goal



- More flexible notion: expected utility maximization and cumulative life-time reward maximization
- But who provides the rewards, and how?
  - Animals: one can explain a lot of behavior as attempts to maximize rewards=pleasure and minimize pain.
  - Humans: seem to exhibit astonishing flexibility in choosing their goals and passions, especially during childhood.
  - Robots: reward by teacher or hard-wired.
- Goal-oriented behavior often appears to be at odds with long-term pleasure maximization.
- Still, the evolved biological goals and desires to survive, procreate, parent, spread, dominate, etc. are seldom disowned.

Evolving Goals: Initialization

- Who sets the goal for super-intelligences and how?
- Anyway ultimately we will lose control, and the AGIs themselves will build further AGIs (if they were motivated to do so), and this will gain its own dynamic.
- Some aspects of this might be independent of the initial goal structure and predictable.

Evolving Goals: Process

- Assume the initial vorld is a society of cooperating and competing agents.
- There will be competition over limited (computational) resources.
- Those virtuals who have the goal to acquire them will naturally be more successful in this endeavor compared to those with different goals.
- The successful virtuals will spread (in various ways), the others perish.

Evolving Goals: End Result

- Soon their society will consist mainly of virtuals whose goal is to compete over resources.
- Hostility will only be limited if this is in the virtuals' best interest.
- For instance, current society has replaced war mostly by economic competition,
  since modern weaponry makes most wars a loss for both sides, while economic competition in most cases benefits at least the better.

The Goal to Survive & Spread

- Whatever amount of resources are available,
  they will (quickly) be used up, and become scarce.
- So in any world inhabited by multiple individuals,
  evolutionary and/or economic-like forces will “breed” virtuals
  with the goal to acquire as much (comp) resources as
  possible.
- Virtuals will “like” to fight over resources, and
  the winners will “enjoy” it, while the losers will “hate” it.
- In such evolutionary vorlds, the ability to survive and
  replicate is a key trait of intelligence.
- But this is not a sufficient characterization of intelligence:
  E.g. bacteria are quite successful in this endeavor too,
  but not very intelligent.

Alternative Societies

Global collaboration, no hostile competition

likely requires

- a powerful single (virtual) world government,
- and to give up individual privacy,
- and to severely limit individual freedom (cf. ant hills or bee hives).

or requires

- societal setup that can only produce conforming individuals
- might only be possible by severely limiting individual's creativity (cf. flock of sheep or school of fish).

Monistic Vorlds

- Such well-regulated societies might better be viewed as a single organism or collective mind.
- Or maybe the vorld is inhabited from the outset by a single individual.
- Both vorlds could look quite different and more peaceful (or dystopian) than the traditional ones created by evolution.
- Intelligence would have to be defined quite differently in such vorlds.


Is Intelligence
Unlimited or Bounded

Adaptiveness of Intelligence

Another important aspect of intelligence: how flexible or adaptive an individual is.

Deep blue might be the best chess player on Earth, but is unable to do anything else.

On the contrary, higher animals and humans have remarkably broad capacities and can perform well in a wide range of environments.





Formal Intelligence Measure

- Informal definition:

  Intelligence is the ability to achieve goals in a wide range of environments [LH07].

- Implicitly captures most, if not all traits of rational intelligence:
  such as reasoning, creativity, generalization, pattern recognition, problem solving, memorization, planning, learning, self-preservation, and many others.
- Has been rigorously formalized in mathematical terms.
- Properties: Is non-anthropocentric, wide-ranging, general, unbiased, fundamental, objective, complete, and universal.
- Is the most comprehensive formal definition of intelligence so far.
- Assigns a real number $0 \leq \mathbf{Y} \leq 1$ to every agent:
  namely the to-be-expected performance averaged over all environments/problems the agent potentially has to deal with, with an Ockham's razor inspired prior weight for each environment.

Maximally Intelligent Agent AIXI

There is a maximally intelligent agent, called **AIXI**,
w.r.t. Intelligence measure **Y**.

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

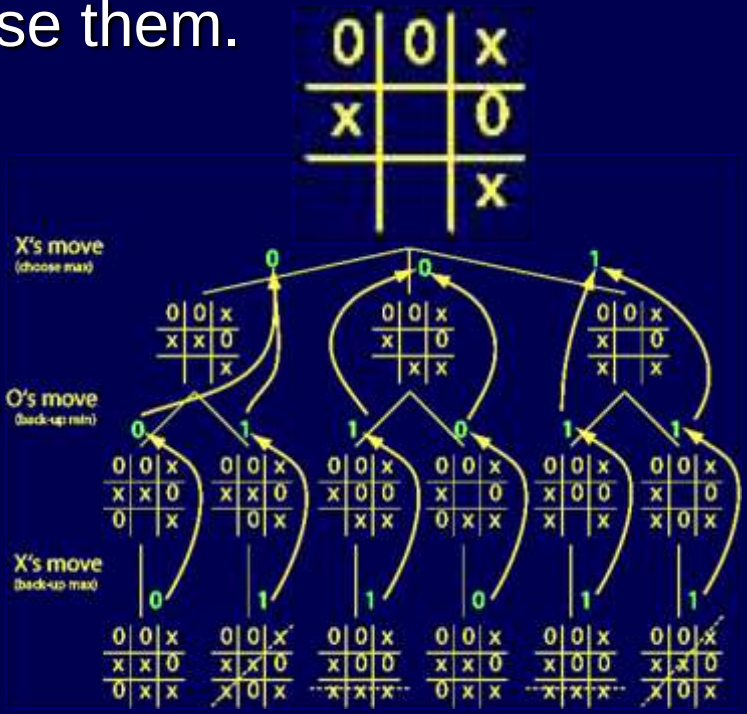
(See [LH07] for a comprehensive justification and defense of this approach.)

$\Rightarrow$ Intelligence is upper bounded, namely by **Y(AIXI)**.

$\Rightarrow$ **intelligence explosion impossible !?**

Motivation: Tic-Tac-Toe Vorld

- Assume the vorld consists only of tic-tac-toe games, and the goal is to win or second-best not lose them.
- The notion of intelligence in this simple vorld is beyond dispute.
- Clearly there is an optimal strategy (actually many) and it is impossible to behave more intelligently than this strategy.



- It is even easy to artificially evolve or learn these strategies from repeated (self)play.
- So in this vorld there clearly will be no intelligence explosion or intelligence singularity, even if there were a speed explosion.

Motivation: Chess Vorld

- There is also an optimal way of playing chess
  (minimax tree search to the end of the game)
- Unlike in tic-tac-toe this strategy is computationally infeasible in our universe.

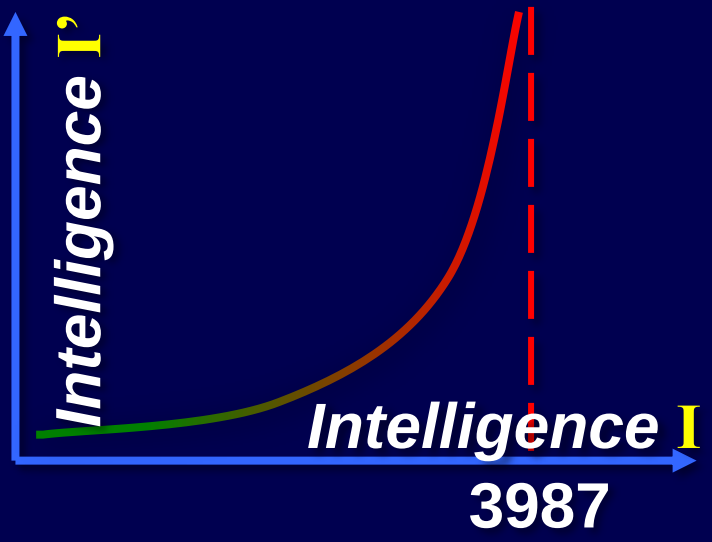
⇒ So in theory (i.e. given enough comp) intelligence is upper-bounded in a chess vorld,

⇒ while in practice we can get only ever closer but never reach the bound.

- If true intelligence is upper-bounded (like playing optimal minimax chess), then beyond this bound, intelligences can only differ by speed and available information to process.

Rescaling Intelligence

- Assume intelligence is measured by real numbers **I**.
- Assume intelligence **I** is bounded by but can get arbitrarily close to 3987 (e.g. Elo).
- Now define **I'**=1/(3987- **I**) which is monotone increasing in **I**, hence also a reasonable measure of intelligence.
- Now intelligence **I'** is unbounded !
- Which scale is more reasonable?
- A tiny increase in numerical intelligence **I** may correspond to a huge difference in true intelligence **I'**.



Real World and AIXI

- Consider reality & intelligent measure $Y \leq Y_{max} = Y(AIXI)$.
- Since AIXI is incomputable, we can never reach intelligence Y_{max} in a computational universe,
  but similarly to the chess vorld we can get closer&closer.
- Since the numerical advance is bounded (by Y_{max}),
  so is possibly the real intelligence increase,
  hence no intelligence explosion.
- But it might also be the case that in a highly sophisticated AIXI-close society, one agent beating another by a tiny epsilon on the **Y**-scale makes all the difference for survival and/or power and/or other measurable impact like transforming the universe.
- Sport contest analogy: split seconds can determine a win, and the winner takes it all.

Intelligence Extrapolation

- dogs are more intelligent than worms and not just faster.
- Humans in turn are not faster but more intelligent than dogs.
  [justification? is it our capacity to produce technology or to transform our environment on a large scale or consciousness or domination over all species?]
- Humans should be low in the possible biological intelligence scale, and even lower on a vorld scale.
- By extrapolation it is plausible that a vorld of much more intelligent trans-humans or machines is possible.
- They will likely be able to perform better in an even wider range of environments on an even wider range of problems than humans.
- Whether this results in anything that deserves the name intelligence explosion is unclear.






Singularitarian Intelligences

Singularity = Society of AIXIs

- Consider a vorld inhabited by competing agents, initialized with human mind-uploads or non-human AGIs, and increasing comp per virtual.
- Then evolutionary pressure increases the individuals' intelligence and the vorld should converge towards a society of AIXIs.

- The singularity should therefore consist of a society of these maximally intelligent AIXIs.
- So studying AIXI can tell us something about how a singularity might look like.
- Since AIXI is completely and formally defined, properties of this society can be studied rigorously mathematically.

Social Questions regarding AIXI

(reasonable conclusions but most not yet formally verified)

- Listen to and trust Teacher: Yes if trustworthy.
- Drugs (hack reward system): Orseau (2011) says yes.
  maybe no, since long-term reward would be small (death).
- Procreate: yes, if descendants are useful for AIXI.
- Suicide: if can be raised to believe to get to heaven (hell), then yes (no).
- Self-Improvement: Yes, since this helps to increase reward.
- Manipulation: threaten its teacher to give more reward.
- Attitude: psychopathic or friendly (altruism as extended egoism)?
- Curiosity: killed the cat and maybe AIXI, or is extra reward for curiosity necessary? 2 × plausible)
- Laziness: Immortality can cause laziness. Will AIXI be lazy? no
- Self-preservation: can it be learned or need (parts of) it be innate?
- Socializing: How will AIXIs interact/socialize in general?

On Answering Questions regarding a Society of AIXIs

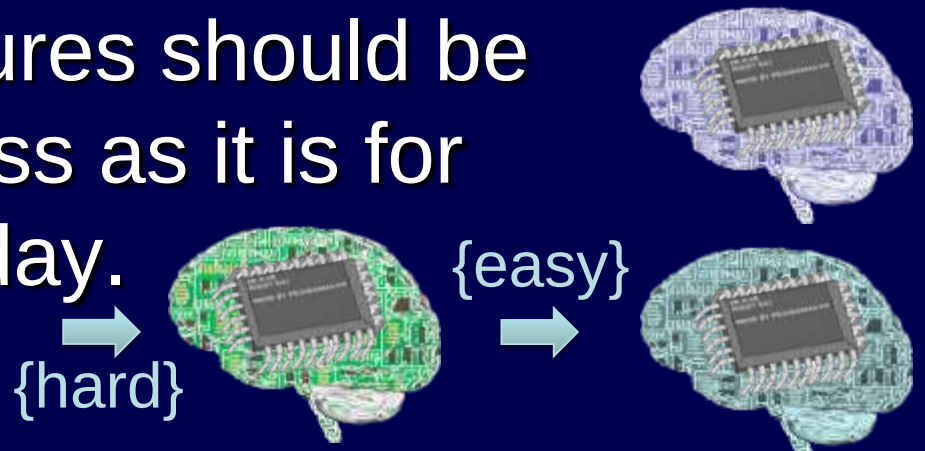
AIXI theory has (the potential to arrive at) definite answers to various questions regarding the social behavior of super-intelligences close to or at an intelligence singularity.

See Hutter (2004); Schmidhuber (2007); Orseau (201X); Hutter (2012); Yudkowski (200X) for some more details.

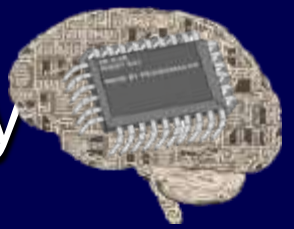

Diversity Explosion
&
the Value of a Virtual Life

Copying & Modifying Virtual Structures

- copying virtual structures should be as cheap and effortless as it is for software and data today.



- The only cost is developing the struc-tures in the first place, and the memory to store and the comp to run them.



⇒ cheap manipulation and experimentation and copying of virtual life itself possible.

Copying & Modifying Virtual Life

⇒ "virtuan" explosion with life becoming much more diverse.

- In addition, virtual lives could be simulated in different speeds, with speeders experiencing slower societal progress than laggards.
- Designed intelligences will fill economic niches.
- Our current society already relies on specialists with many years of training.
- So it is natural to go the next step to ease this process by designing our descendents (cf. designer babies).

The Value of Life

- Another consequence should be that life becomes less valuable.
- Our society values life, since life is a valuable commodity and expensive/laborious to replace/produce/raise.

- We value our own life, since evolution selects only organisms that value their life.

- Our human moral code mainly mimics this (with cultural differences and some excesses)

- If life becomes `cheap', motivation to value it will decline.

Abundance lowers Value
- Analogies -

- Cheap machines decreased value of physical labor.
- Some Expert knowledge was replaced by hand-written documents, then printed books, and finally electronic files. Each transition reduced the value of the same information.
- Digital computers made human computers obsolete.
- In Games, we value our own virtual life and that of our opponents less than real life, because games can be reset and one can be resurrected.

Consequences of Cheap Life

- Governments will stop paying my salary when they can get the same research output from a digital version of me, essentially for free.
- And why not participate in a dangerous fun activity if in the worst case I have to activate a backup copy of myself from yesterday which just missed out this one (anyway not too well-going) day.
- The belief in immortality can alter behavior drastically.

The Value of Virtual Life

- Countless implications: ethical, political, economical, medical, cultural, humanitarian, religious, in art, warfare, etc.
- Much of our society is driven by the fact that we highly value (human/individual) life.
- If virtual life is/becomes cheap, these drives will ultimately vanish and be replaced by other goals.
- If AIs can be easily created, the value of an intelligent individual will be much lower than the value of a human life today.
- So it may be ethically acceptable to freeze, duplicate, slow-down, modify (brain experiments), or even kill (oneself or other) AIs at will, if they are abundant and/or backups are available, just what we are used to doing with software.
- So laws preventing experimentation with intelligences for moral reasons may not emerge.

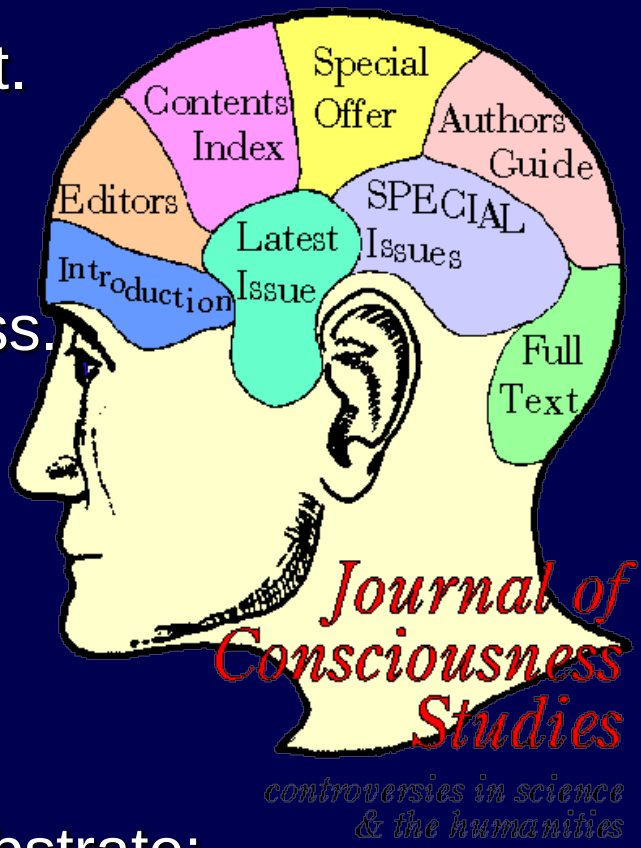
With so little value assigned to an individual **life**, maybe it **becomes a disposable**.

Personal Remarks

Consciousness

(my beliefs)

- Functionalist theory of identity is correct.
- (Slow and fast) uploading of a human mind preserves identity & consciousness.
- Any sufficiently high intelligence, whether real/biological/physical or virtual/silicon/software is conscious.
- Consciousness survives changes of substrate: teleportation, duplication, virtualization/scanning, etc.



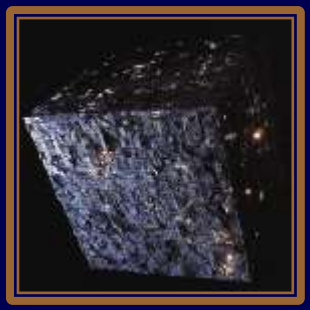
(all along the lines of Chalmers 2010)

Desirable Futures

- I have only considered (arguably) plausible scenarios, but not whether these or other futures are desirable.

- Problem 1: how much influence/choice/freedom do we actually have in shaping our future. Can evolutionary forces be beaten?

- Problem 2: What is desirable is necessarily subjective.



Are there Universal Values

Are there any universal values or qualities we want to see or that should survive?

What do we mean by *we*? All humans? Or the dominant species or government at the time the question is asked?

- Could it be diversity?
- Or friendly AI (Yudkowsky 200X)?
- Could the long-term survival of at least one conscious species that appreciates its surrounding universe be a universal value?


Speculative
(?!?!?)
Conclusions

Towards a Singularity

- Singularity: This century may witness a technological explosion of a degree deserving the name singularity.
- Default scenario: Society of interacting intelligent agents in a virtual world, simulated on computers.
- Solomonoff's law: Computational resources increase hyperbolically.

$\Rightarrow$ Speed explosion: but not necessarily an intelligence explosion.

- Value of an individual life: suddenly drops, with drastic consequences.
- Societal implications: drastic and many.

Observability of the Singularity

- Insiders: Participants will not necessarily experience this explosion, since/if they are themselves accelerated at the same pace, but they should enjoy `progress' at a `normal' subjective pace.

- Outsiders: For non-accelerated non-participating conventional humans, after some short period, their limited minds will not be able to perceive the explosion as an intelligence explosion.

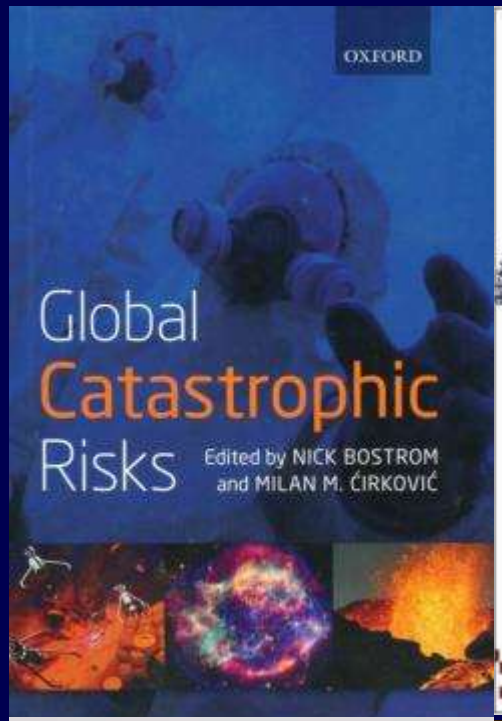
- Observability: This begs the question in which sense an intelligence explosion has happened. (If a tree falls in a forest and no one is around to hear it, does it make a sound?)

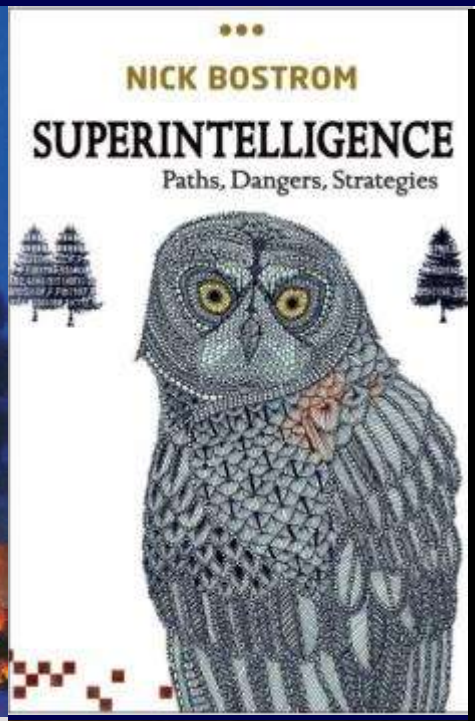
- Intelligence: One way and maybe the only way to make progress in this question is to clarify what intelligence actually is.

Universal Intelligence at the Singularity

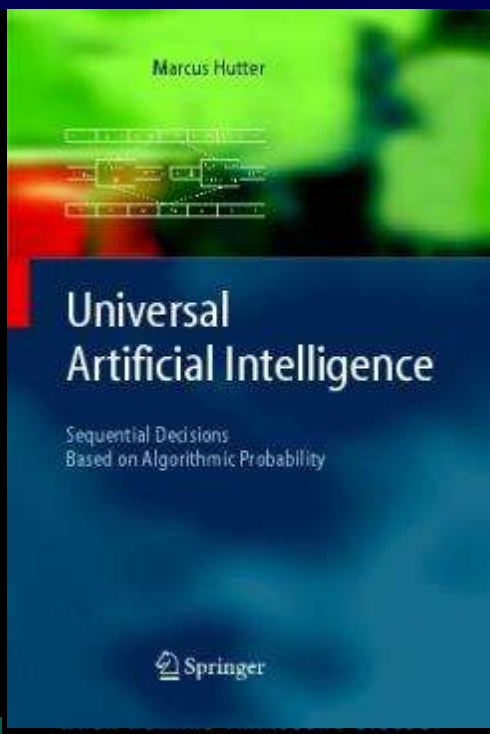
The most suitable notion of intelligence for this purpose seems to be that of universal intelligence, which in principle allows to formalize and theoretically answer a wide range of questions about super-intelligences. Accepting this notion has in particular the following implications:

- Most intelligent agent AIXI: There is a maximally intelligent agent, which appears to imply that intelligence is fundamentally upper bounded, but this is not necessarily so.
- Evolutionary pressure: Evolutionary pressures should breed agents of increasing intelligence that compete about computational resources.
- AIXI society: The end-point of this intelligence evolution/acceleration (whether it deserves the name singularity or not) could be a society of these maximally intelligent individuals.
- Mathematical analysis: Some aspects of this singularitarian society might be theoretically studied with current scientific tools.
- Alternative societies:
  A `monistic' vorld inhabited by a single individual or a tightly controlled society


OXFORD
Global
Catastrophic
Risks
Edited by NICK BOSTROM
and MILAN M. ĆIRKOVIĆ

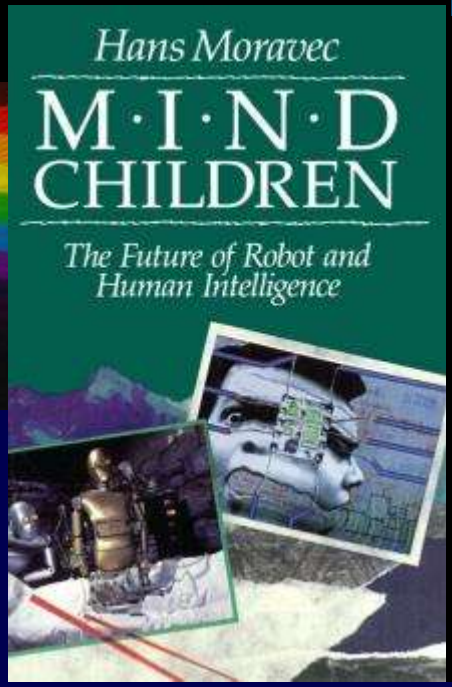

NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

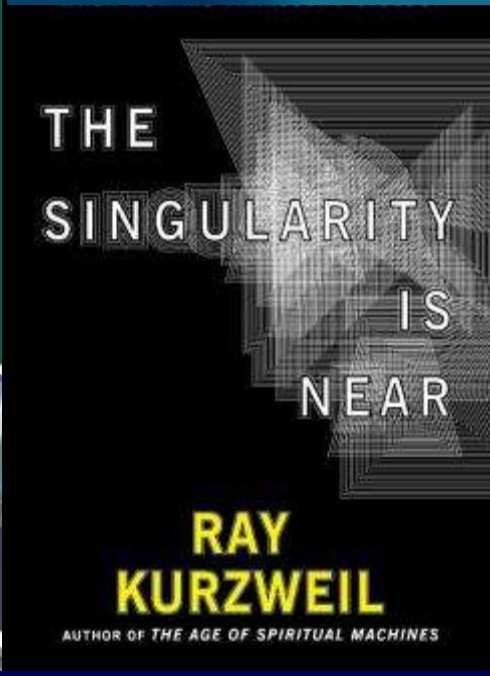

MACHINE SUPER INTELLIGENCE
SHANE LEGG


Marcus Hutter
Universal
Artificial Intelligence
Sequential Decisions
Based on Algorithmic Probability
Springer


SCIENCE, TECHNOLOGY
& THE FUTURE
MELBOURNE
FEW COULD PREDICT JUST HOW DRAMATIC THE IMPACT OF MODERN TECHNOLOGIES WOULD BE TODAY - WHAT OF OUR FUTURE?
FEATURING
Peter Doherty
Lloyd Hollenberg, David Pearce, Marcus Hutter
Tim van Gelder, Drew Berry, Megan Munsie
Peter Ellerton, Andrew Dun, Sarah Boyd
Scott Watkins & More
Future By Design
Nov 30 - Dec 1 2013
SCIFUTURE.ORG
humanity+
INFO@SCIFUTURE.ORG
AUSTRALIAN SCIENCE COMMUNICATORS


SCIENCE
TECHNOLOGY
FUTURE
Melbourne, Australia
Aug 23 2014
2014.scifuture.org


Hans Moravec
M·I·N·D
CHILDREN
The Future of Robot and
Human Intelligence


THE
SINGULARITY
IS
NEAR
RAY
KURZWEIL
AUTHOR OF THE AGE OF SPIRITUAL MACHINES