Implications of Artificial Intelligence

– the most transformative technology of the 21st century

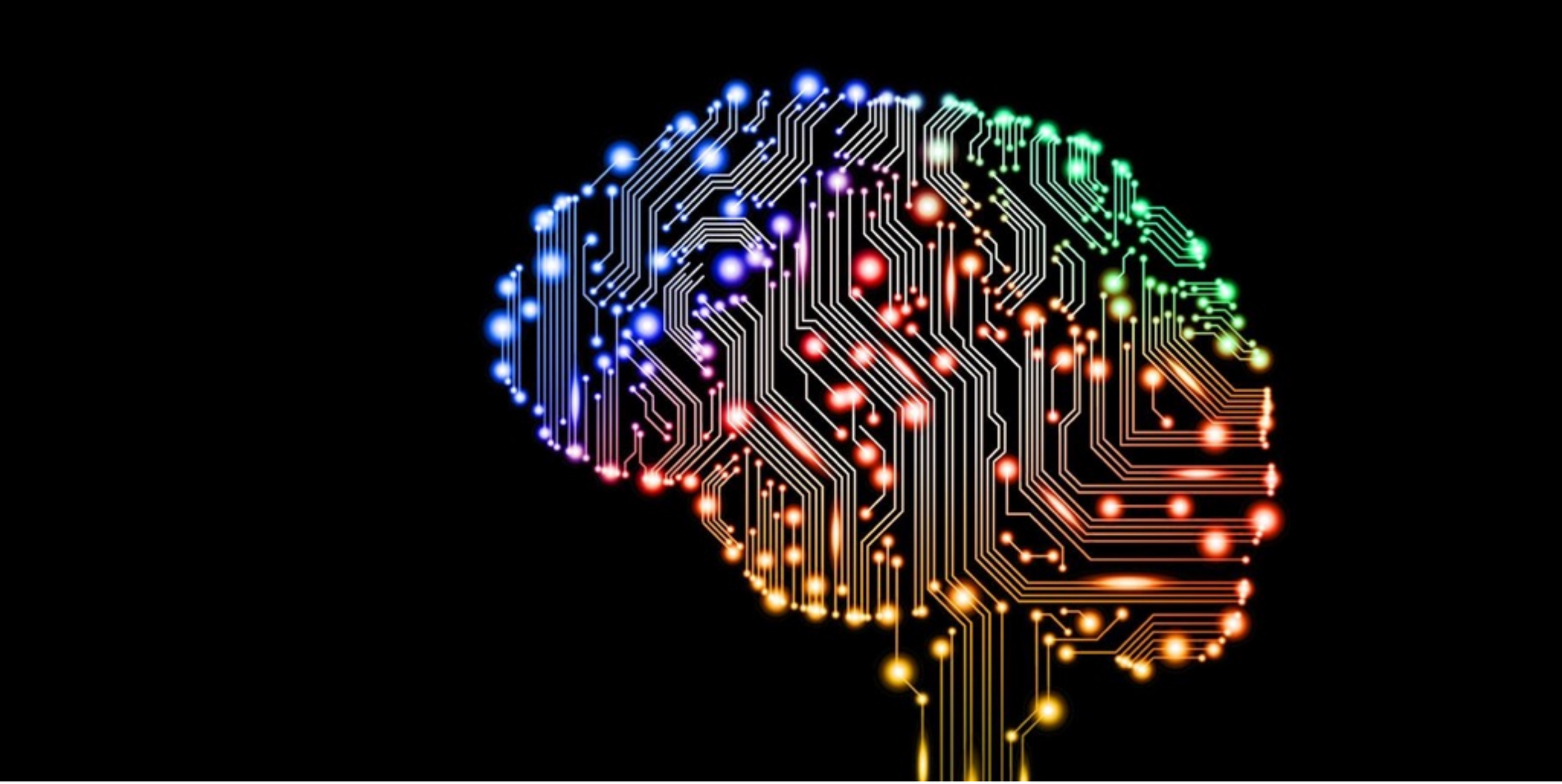
STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

Kaspar Etter
www.superintelligence.ch

Robin Li, Bill Gates, Elon Musk on AI

More Money for Entertainment

… than ensuring a good outcome!



Introduction

What are we talking about?

STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

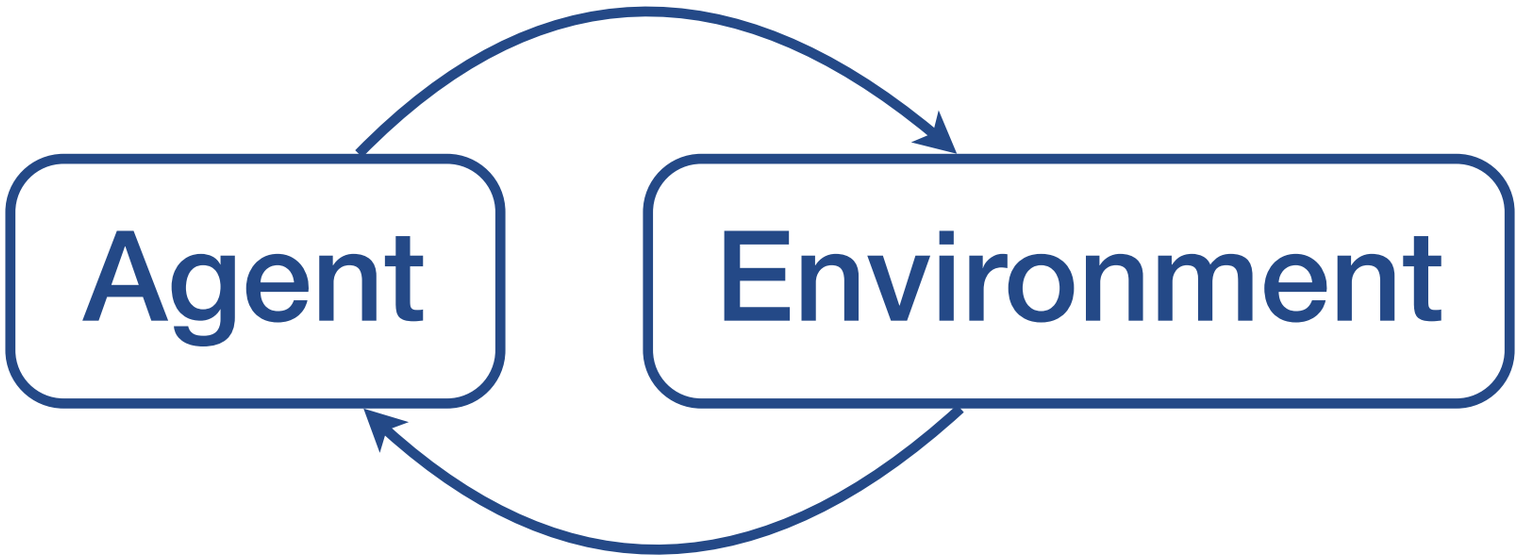
Intelligence

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

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

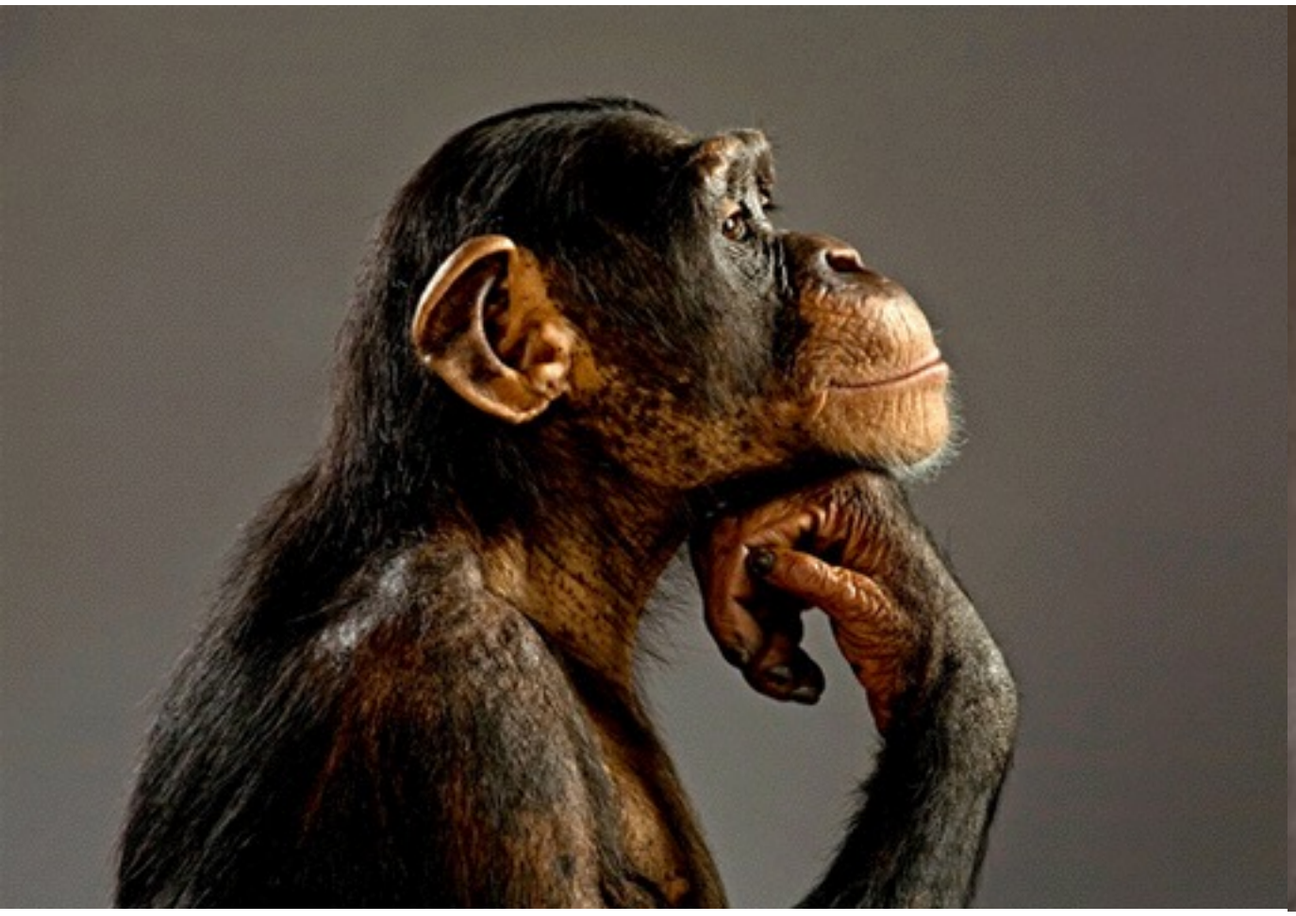
Universal Intelligence
arxiv.org/pdf/0712.3329.pdf

Ingredients



Learn, predict, rate and plan!

Intelligence is a Big Deal

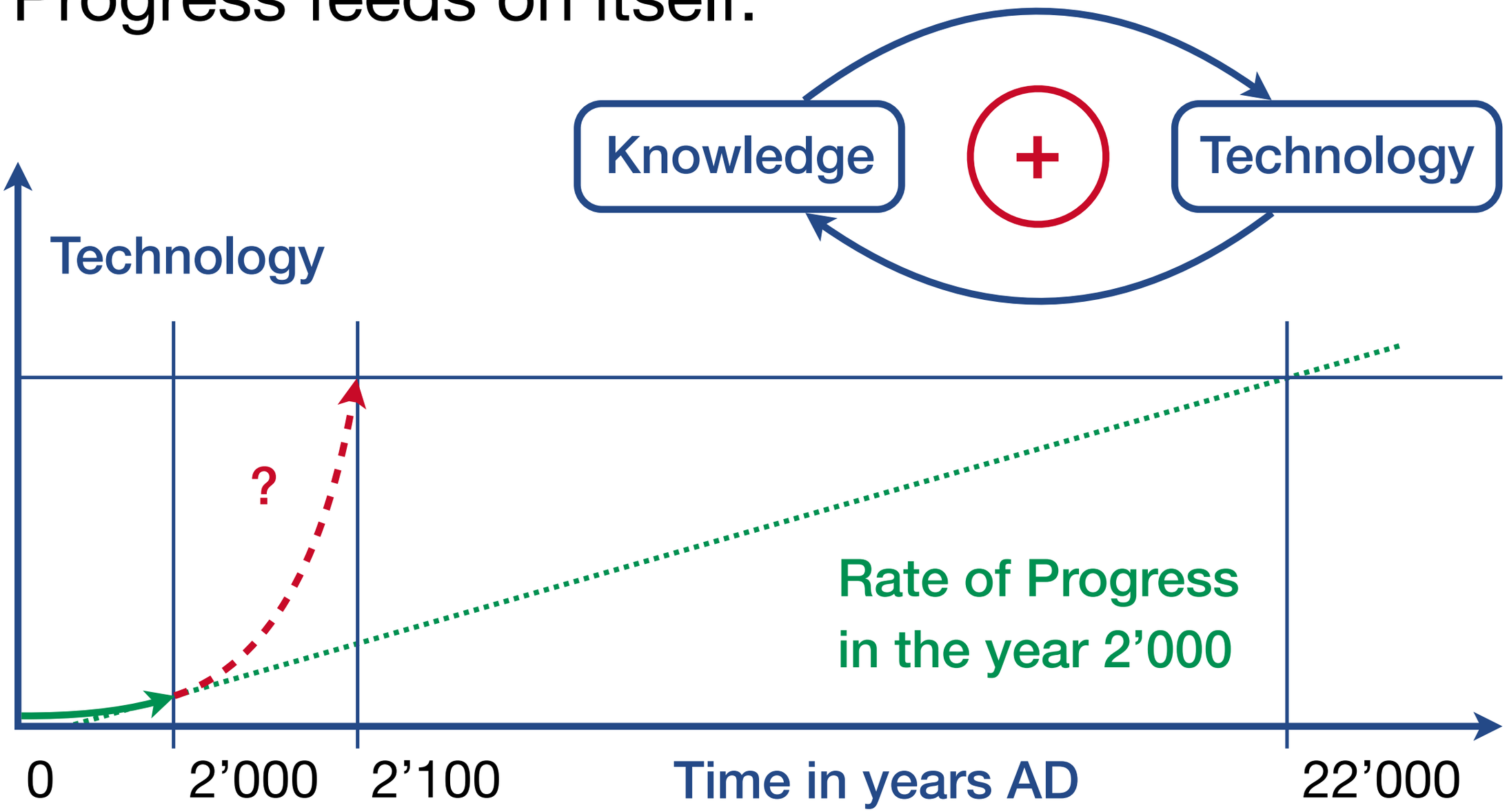


6 million years ago, 96% common DNA

Fast-Evolving Human DNA Leads to Bigger-Brained Mice
phenomena.nationalgeographic.com/2015/02/19/[…]

Accelerating Change

Progress feeds on itself:



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

Technology = (Neutral) Lever

Greed and laziness drive us to

– increase our **productivity**

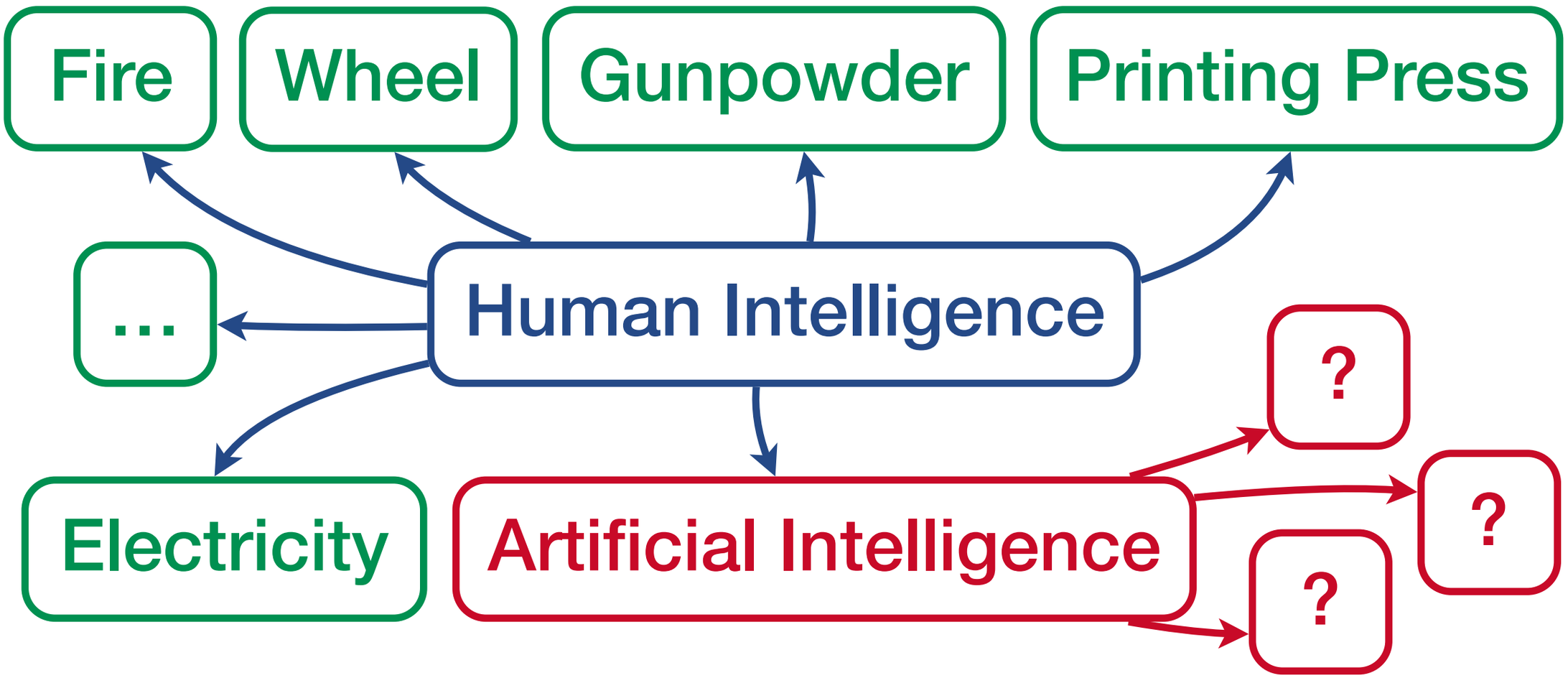
– make our **lives easier**

… which is fine except:

Our technological progress far outperforms our moral progress!

Dual-Use Technology
en.wikipedia.org/wiki/Dual-use_technology

Artificial Intelligence



Intelligence is a technology like no other!

Current Trends

Where are we heading to?

STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

State of the Art

Deep Blue: 1997

Stanley: 2005

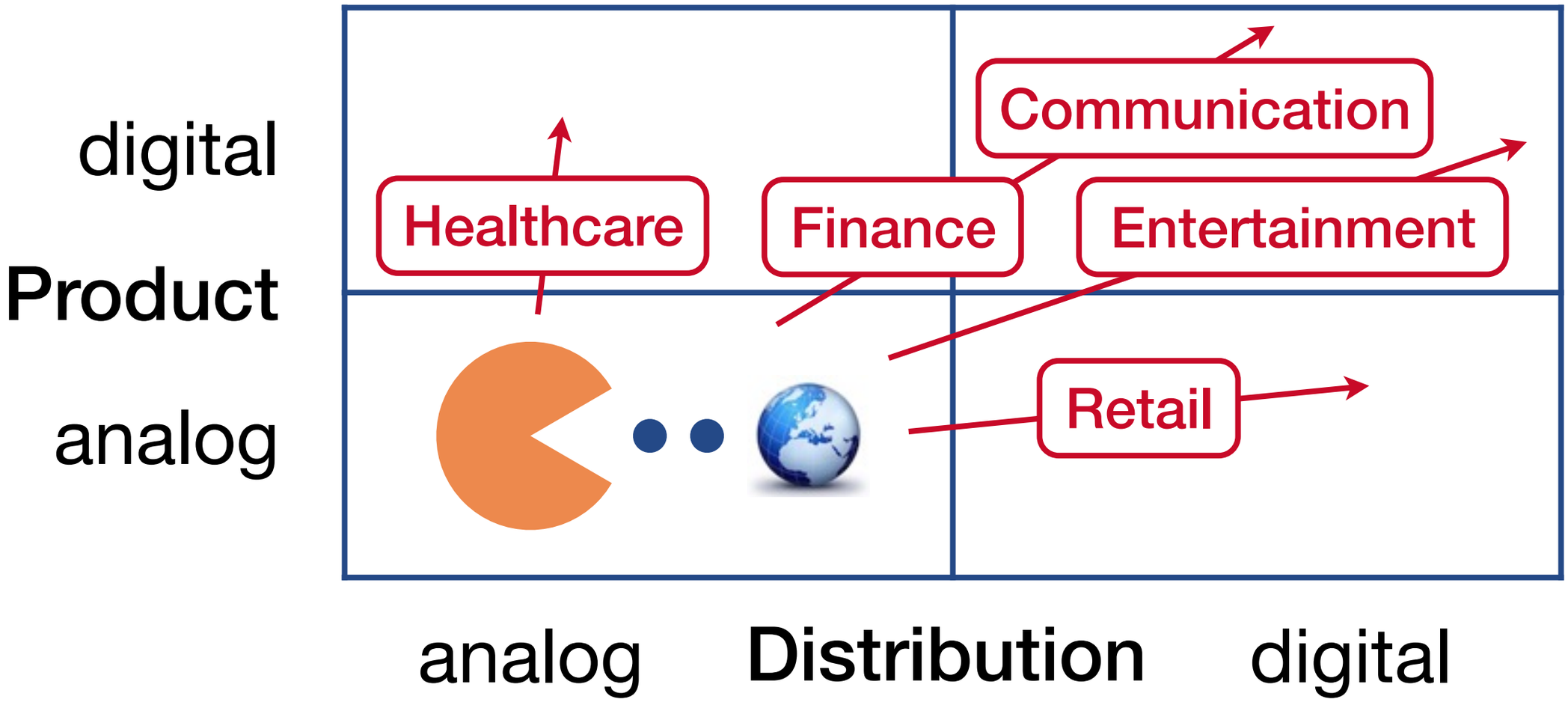
IBM Watson: 2011

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	Superhuman

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

Software is eating the world



… and AI will digest all this (big) data!

Marc Andreessen
on.wsj.com/1w2FbVs

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

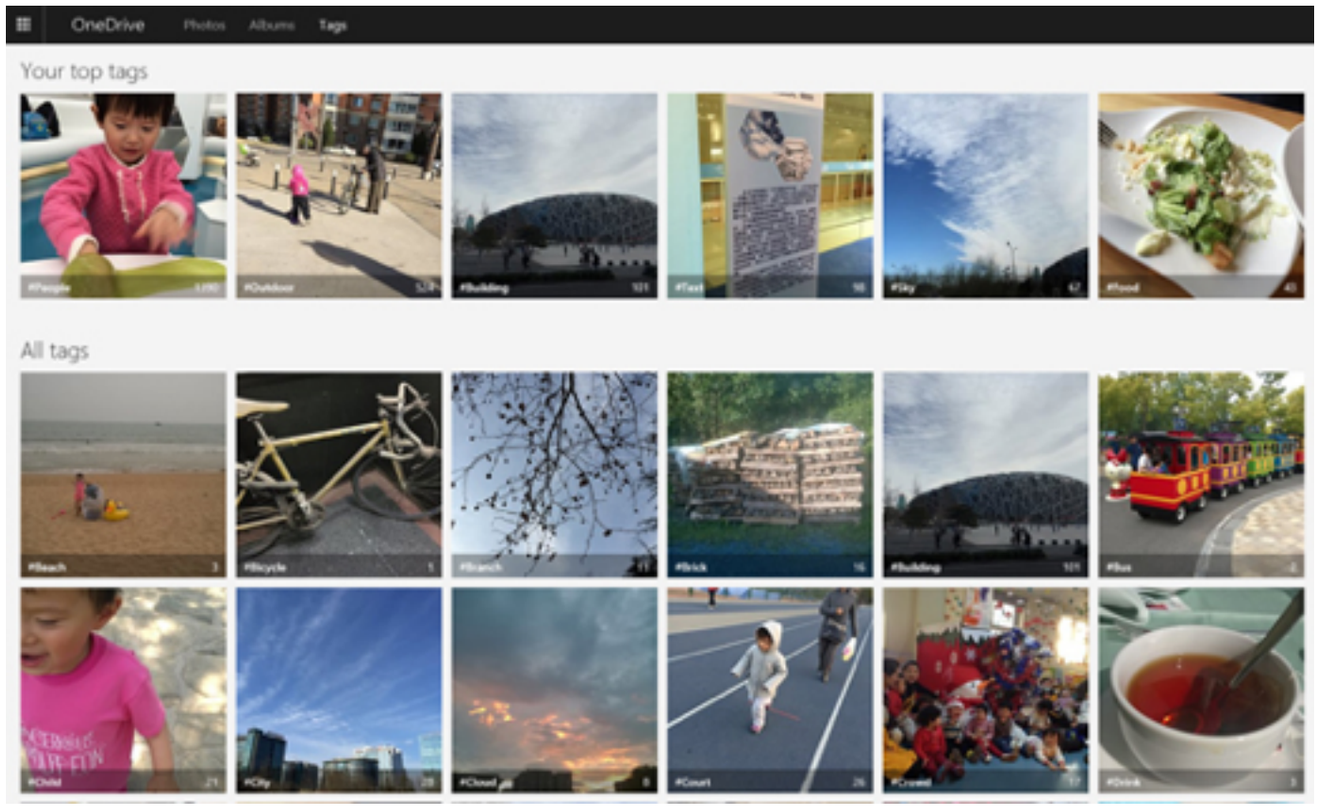
A yellow school bus parked in a parking lot.

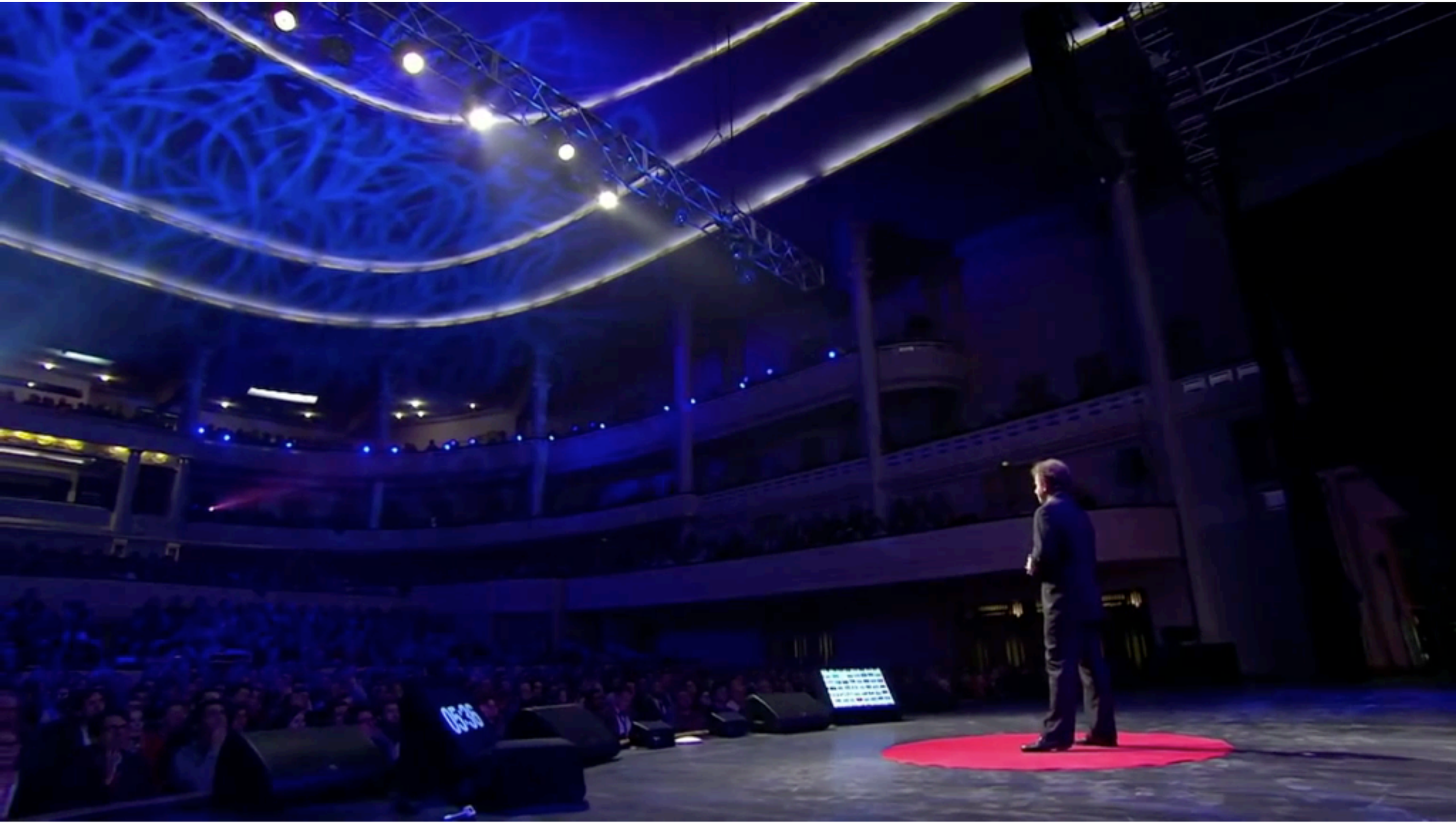
Google Research Blog, November 2014
googleresearch.blogspot.ch/2014/11/[…]

Superhuman Image Recognition

– With convolutional neural networks

– 1.2 m training images, ~ 30 layers





Jeremy Howard
go.ted.com/bbZC

MACHINE INTELLIGENCE 2.0

AGENTS

PROFESSIONAL
Howdy! · x.ai · clara · KASISTO · DigitalGenius · OVERLAP.CC · meekan · fuse|machines · PRIMER

PERSONAL
facebook · XIAOICE · assistant.ai · large · nestor · @wesome · Magic

OS INTERFACES
Siri · Cortana · VIV · maluuba · api.ai · CogNEA · Google Now

AUTONOMOUS SYSTEMS

AIR
3DR · dji · PROJECT LOON (Google) · VERTICAL · DroneDeploy · AIRDOG · SKYCATCH · SKYDIO · Airware · LILY

GROUND
Google · UBER · TESLA · CRUISE · MOBILEYE · COMMA · AdasWorks

SEA
LIQUID ROBOTICS · bluefin data · OPENROV · BluHaptics

INDUSTRIAL
KIVA Systems · fetch · HARVEST · CLEARPATH · AVIDBOTS · ENERGID · rethink robotics · GREYORANGE · OSARO

ENTERPRISE

SECURITY / FRAUD
Sentinel · graphistry · BITSIGHT · feedzai · AREA 1 · drawbridge · sift science · CYLANCE · Brighterion

HR / RECRUITING
textio · hiQ · gild · SpringRole · entelo · unitive · GIGSTER

SALES
6sense · clari · infer · Preact · Gainsight · AVISO · salespredict · sentient · Vidora · people pattern · Prism

MARKETING
LiftIgniter · RADIUS · brightfunnel · retention SCIENCE · AIRPR

CUSTOMER SUPPORT
CLARABRIDGE · QUANTIFIND · Wisedo · ACTIONIQ · FRAMED · DigitalGenius

INTERNAL INTEL
Alation · ADATAD · Palantir · sapho · lucid · Rainbird · SKIPFLAG · algolo · Digital Reasoning · Narrative Science

MARKET INTEL
Quid · mattermark · DataFox · bottlenose · PREMISE · enigma · CB INSIGHTS

PLATFORMS

RESEARCH / AGI
OpenAI · vicarious · Google DeepMind · Numenta · Cycorp · nnaisense · SCALED INFERENCE · CURIOUS · GEOMETRIC INTELLIGENCE

FULL STACK
context relevant · CognitiveScale · NVIDIA · TERADEEP · QUALCOMM · nervana SYSTEMS

MACHINE LEARNING
Dato · rapidminer · cortical.io · AYASDI · amazon web services · Azure Machine Learning · narrative logics · PredictionIO · SKYTREE · big · blue yonder

INDUSTRIAL IOT
ThingWorx · UPTAKE · IMUBIT · Preferred Networks · Alluvium · xively · PLANET OS

AUDIO
Gridspace · TalkIQ · nexidia · vocaliq · NUANCE · Expect Labs · popup archive

VISION
ORBITAL INSIGHT · Descartes Labs · DEXTRO · cortica · clarifai · PLANET LABS · MetaMind

DATA ENRICHMENT
diffbot · Paxata · TRIFACTA · iDIBON · WorkFusion · loop AI Labs · CrowdFlower

INDUSTRIES

ADTECH
ROTHEORENT · dstillery · BEYONDVERBAL · METAMARKETS · TAPAD · rocketfuel · affectiva

AGRICULTURE
BLUE RIVER · tule · TerrAvion · mavrx · THE CLIMATE CORPORATION · CERES TECHNOLOGIES · HONEYCOMB

FOR GOOD
Conservation Metrics · DataKind · DATA POP · thorn · BAYES IMPACT

RETAIL FINANCE
inVenture · Affirm · earnest · MIRADOR · Lendo · zest finance · LendUp

LEGAL
Everlaw · RAVEL · LEGAL ROBOT · Seal · BEAGLE · ROSS · Lex Machina

MATERIALS & MFG
zymergen · AUGMATE · GINKGO BIOWORKS · ITRINE TECHNOLOGIES · SIGHT MACHINE · CALCULARIO · Eigen Innovations

HEALTHCARE
deep genomics · 3SCAN · enlitic · Calico · Atomwise · Recombine · colcr · METABIOTA · GRAND ROUNDS · Google Life Sciences · IBM Watson Health

INDUSTRIES (CONT'D)

EDUCATION
KNEWTON · coursera · turnitin · gradescope · UDACITY · KHANACADEMY

TRANSPORT & LOGISTICS
NAUTO · taleris · PRETECKT · clearmetal

INVESTMENT FINANCE
Bloomberg · Quantopian · Dataminr · KENSHO · iSENTIUM · NEURENSIC · alphasense

TECH USER TOOLS

DATA SCIENCE
DOMINO · kaggle · Sentenai · sense · yseop · Outlier · ŷhat · DataRobot

MACHINE LEARNING
Cortana Analytics · AlchemyAPI · glowfish · IBM Watson Platform · Anodot · MonkeyLearn · (h[s]) HyperScience · fuzzy.io · SIGOPT · 0xdata H2O · SPARKBEYOND · indico

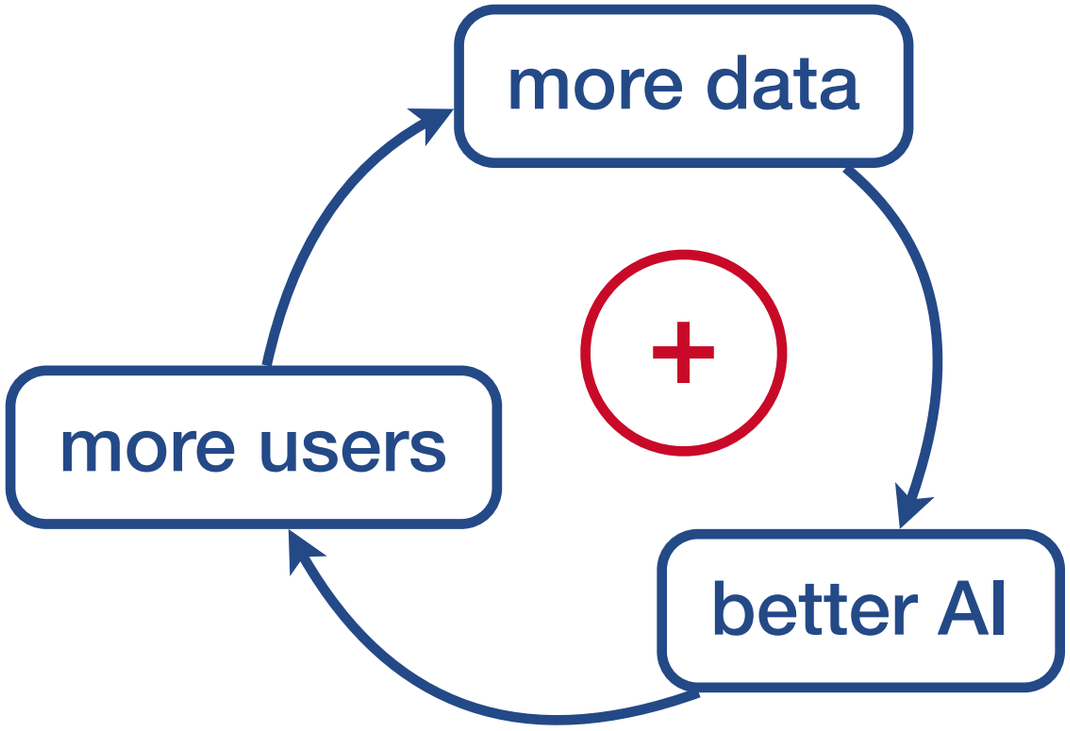
OPEN SOURCE
SKYMIND · TensorFlow · seldon · Caffe · theano · Spark MLlib · Microsoft DMTK · spaCy · DL4J · SciKit · CGT

SHIVONZILIS.COM/MACHINEINTELLIGENCE

Bloomberg BETA

Deep Learning Explosion

Stuart Russell: «Industry [has probably invested] more in the last 5 years than governments have invested since the beginning of the field [1950s].»



Massive economic incentives (and first mover advantages)

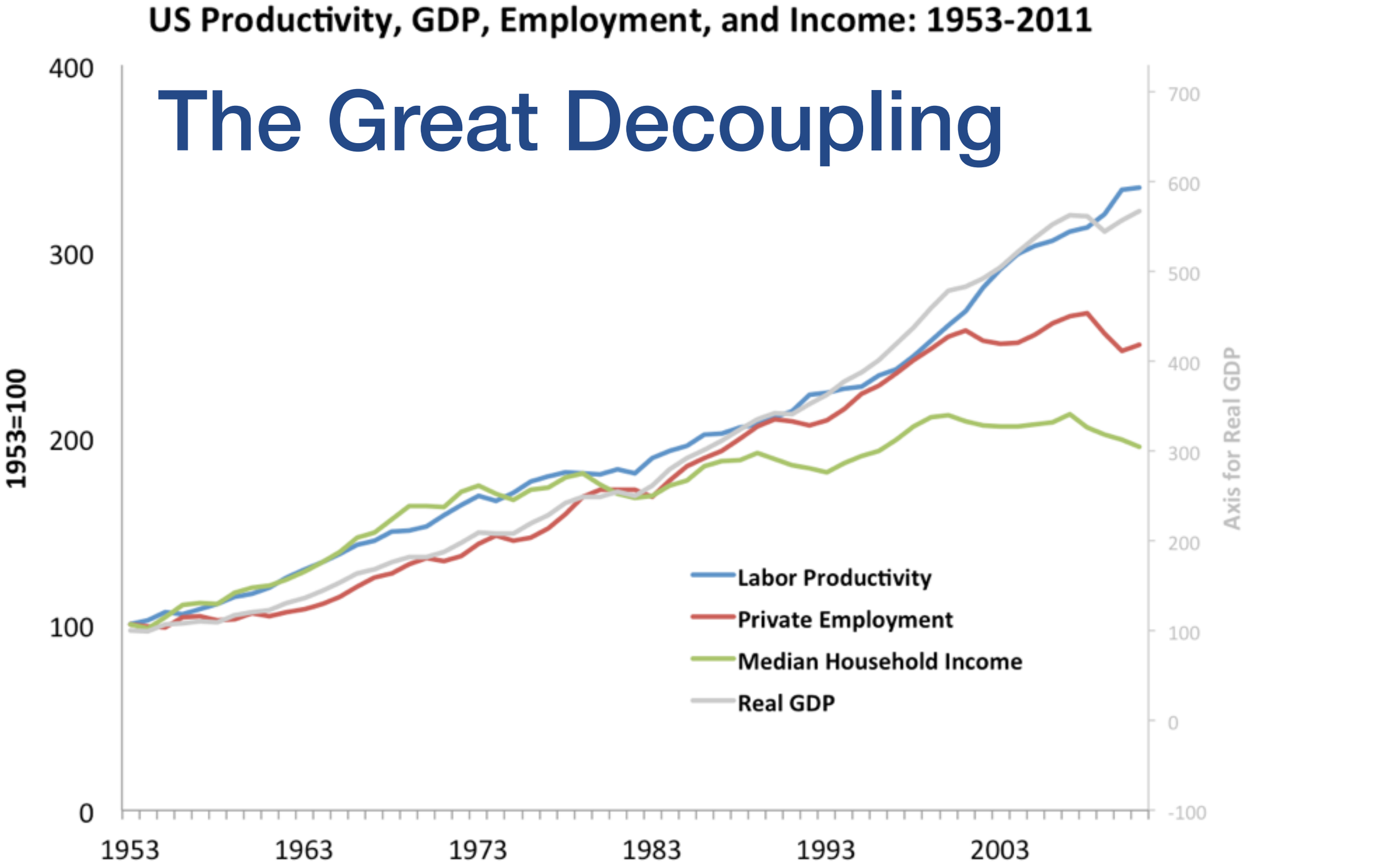
How Big is the Field of Artificial Intelligence?
intelligence.org/2014/01/28/how-big-is-ai/

Automation brings abundance

Many industries will be transformed into pure AI industries…

Capital: ⬆, Wages: ⬇

Should be beneficial if we manage it well – but we are not prepared!



Sources: Census Bureau, Bureau of Labor Statistics

The Great Decoupling of the US Economy
andrewmcafee.org/2012/12/[…]

We'll give robots full autonomy

… because of

- increased speed
- high complexity
- risk of jamming

Current examples:

- financial markets
- auton. weapons

AI is the ultimate productivity boost!

Daniel Suarez
go.ted.com/biXE

Machine Learning by Google

DEEPMIND

Google's Algorithms Learn Tasks Independently
www.theguardian.com/technology/2015/feb/25/[…]

Reinforcement Learning

Train with first-person shooters and deploy on armed drones…

Visual Doom AI Competition

«Can AI effectively play Doom using only raw visual input?»

Visual Doom AI Competition @ CIG 2016
vizdoom.cs.put.edu.pl/competition-cig-2016

Challenges

What might go wrong?

STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

A rational agent will strive to …

- stay functional (**self-preservation**)
- keep its goal (**goal-preservation**)
- get stuff (**resource accumulation**)
- be smarter (**intelligence explosion**)

The problem is not malevolence but different goals and higher decision quality!

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

Single-Shot Situation

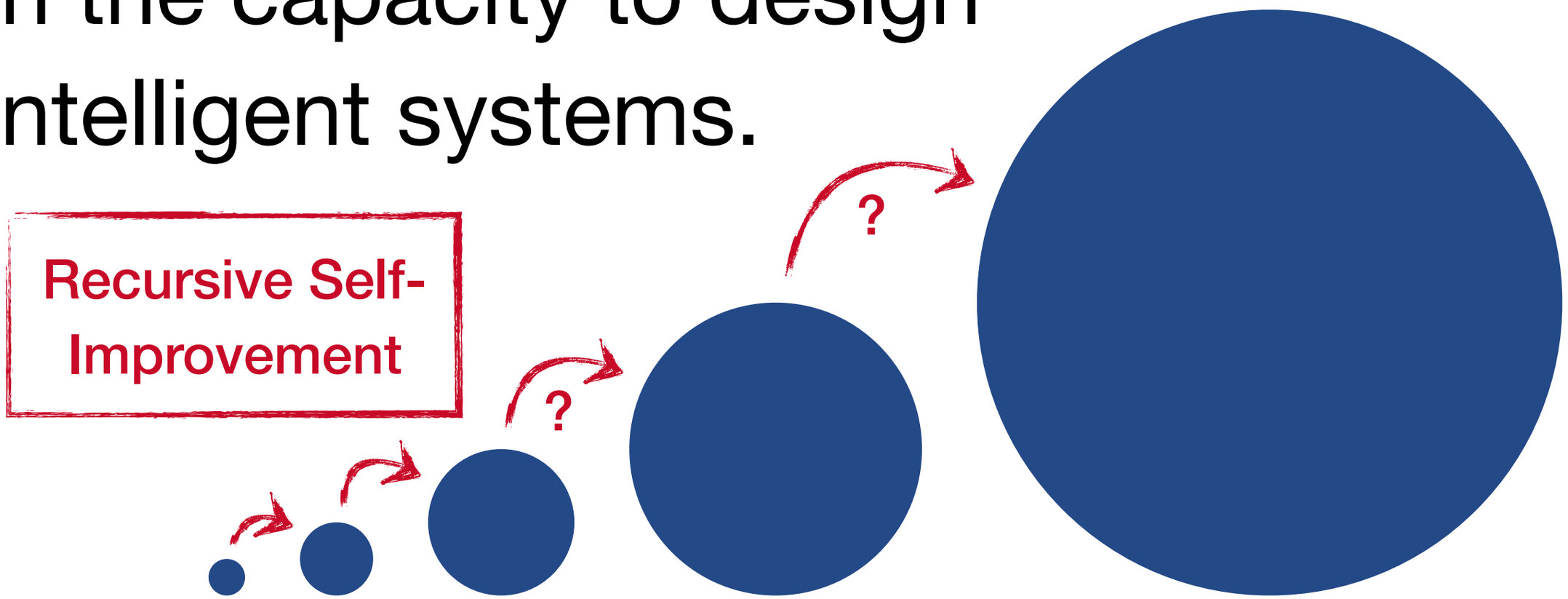
Our first superhuman AI must be a safe one for we may not get a second chance!

- 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

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

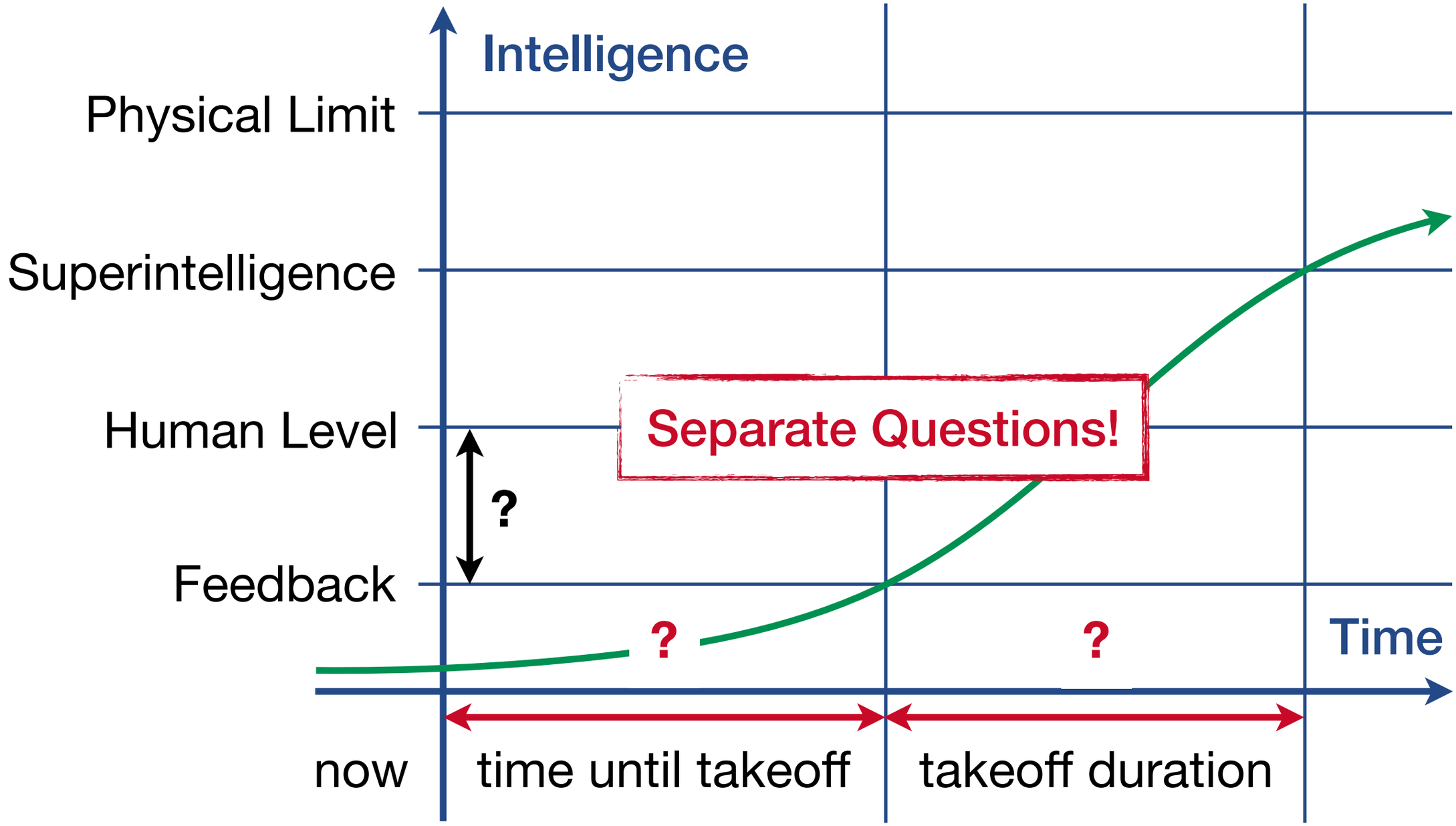
Theoretic phenomenon: There are arguments why it should exist but it has not *yet* been confirmed experimentally.

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

Takeoff Scenarios



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

Interview by John Oliver with Stephen Hawking
www.youtube.com/watch?v=T8y5EXFMD4s

Modelling Capabilities

An advanced AI will also model its operators and go to great lengths to prevent being switched off!

It will behave nicely and cooperatively until the **external threats** are under control and it is ready for takeover.

Ex Machina Movie
exmachina-movie.com

Optimization Power

Problem: When optimizing a system, unspecified parameters often assume extreme values.

You will get what you wished for and not what you wanted

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

Outlook

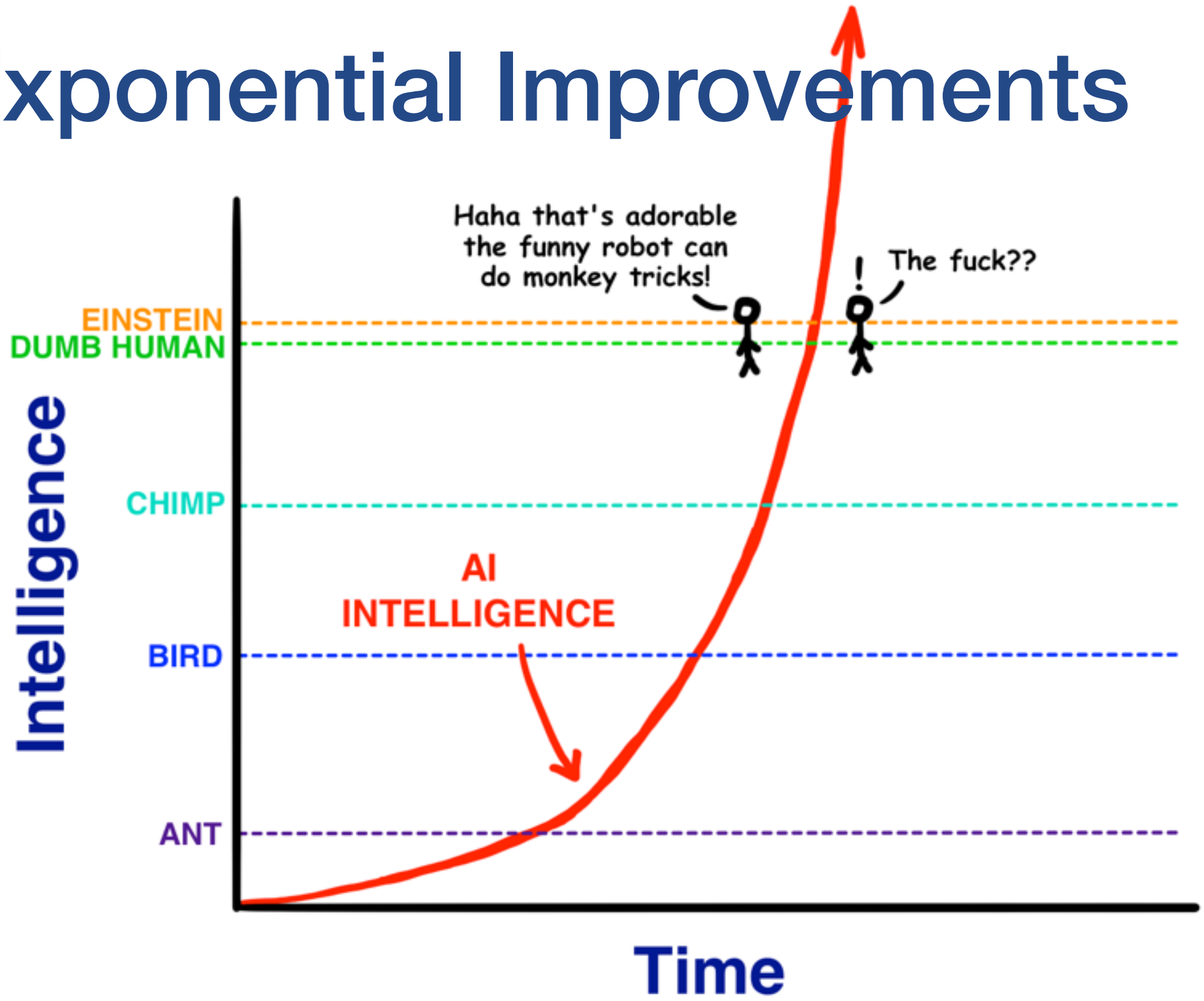
What is ahead of us?

STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

Implications of (advanced) AI

- **Short-term** (now to 15 years): classification/ assistance, surveillance/intelligence, autonomous systems (cars, weapons, etc.) (AIs as moral agents)
- **Mid-term** (15 to 40 years): automation/unemploym., cooperation/arms races, control/principal agent problem, genie/oracle problem (unintended effects)
- **Long-term** (40 to 100 years): scientific breakthroughs by AIs, recursive self-improvement/runaway AI (?), artificial consciousness (?) (AIs as moral patients)

Exponential Improvements



The Road to Superintelligence
waitbutwhy.com/2015/01/[…]

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

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!

The breakthroughs that have unleashed AI on the world
www.wired.com/2014/10/future-of-artificial-intelligence/

Strategy

What is to be done?

STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

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

AI Safety Research

- Value alignment
- Corrigibility
- Security
- Verification
- Transparency
- Other things

Drop me a line at me@kasparetter.com if you're interested to work on these problems! (I will connect you with researchers in the field.)

Research into risks from artificial intelligence
80000hours.org/[…]/artificial-intelligence-risk-research/

Differential Intellectual Progress

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

AI safety should outpace **AI capability** research

Who is working on this?

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

future of life INSTITUTE

Future of Humanity Institute
UNIVERSITY OF OXFORD

CENTRE FOR THE STUDY OF EXISTENTIAL RISK
UNIVERSITY OF CAMBRIDGE

Open Philanthropy Project

Foundational Research INSTITUTE

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!

2014: A turning point in AI safety!

Many smart people take superintelligence very seriously.

Future of Life Institute: AI Conference
futureoflife.org/misc/ai_conference

With great power comes great responsibility!