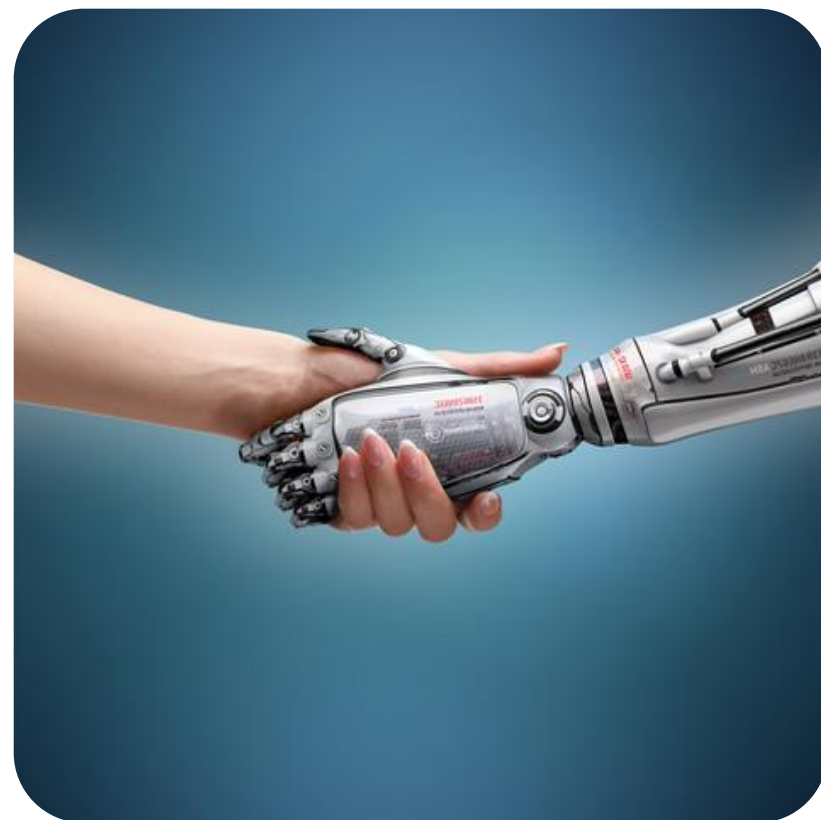


# Artificial Intelligence

## V12: AI & Society

Responsible AI development & deployment  
Unintended threats through AI systems  
Guarding against malicious use  
Possible futures

Using material from Reinhard Karger, DFKI



# Educational objectives

- **Know** the **FAT ML principles** of accountable AI
- **Know** possible **futures** of society as enabled/induced by AI
- **Understand** the **discussions**, expectations and fears about AI in the **general public**
- **Be able to** engage in an educated discussion on the topic, **countering fear**, uncertainty and doubt
- **Act responsibly** in the context of developing and applying AI



# 1. RESPONSIBLE AI DEVELOPMENT & DEPLOYMENT



Source: <https://www.acamstoday.org/dual-use-technology-facilitation/>

# Developing for algorithmic fairness

FAT / ML

## The FAT ML code of conduct

See <http://www.fatml.org/resources/principles-for-accountable-algorithms>

### Purpose

- Help developers to **build algorithmic systems in publicly accountable ways**
- Accountability: the **obligation to report, explain, or justify** algorithmic decision-making & **mitigate** any **negative social impacts** or potential harms

### Premise

- *A **human ultimately responsible** for decisions made/informed by an algorithm*

### Principles

- **Responsibility, Explainability, Accuracy, Auditability, Fairness**

Make available somebody who will take care of adverse individual / societal effects

Explain any **algorithmic decision** in non-technical terms to end users

Report all **sources of uncertainty** / error in algorithms & data

Enable 3<sup>rd</sup> parties to **probe & understand** system **behavior**

Ensure algorithmic **decisions are not discriminatory** w.r.t. to people groups

### Making it actionable

- **Publish a Social Impact Statement**
- ...use above **principles as a guiding structure**
- ...**revisit three times** during development process: design stage, pre-launch, post-launch

## 2. UNINTENDED THREATS THROUGH AI SYSTEMS

**AI System** (definition):

any technical system (software and/or hardware)  
based on digital technology  
that performs tasks *commonly thought* to require intelligence.

# Algorithmic bias

Wikipedia: “**Algorithmic bias** occurs when a computer system behaves in ways that reflects the implicit values of humans involved in the **data collection, selection, or use.**”

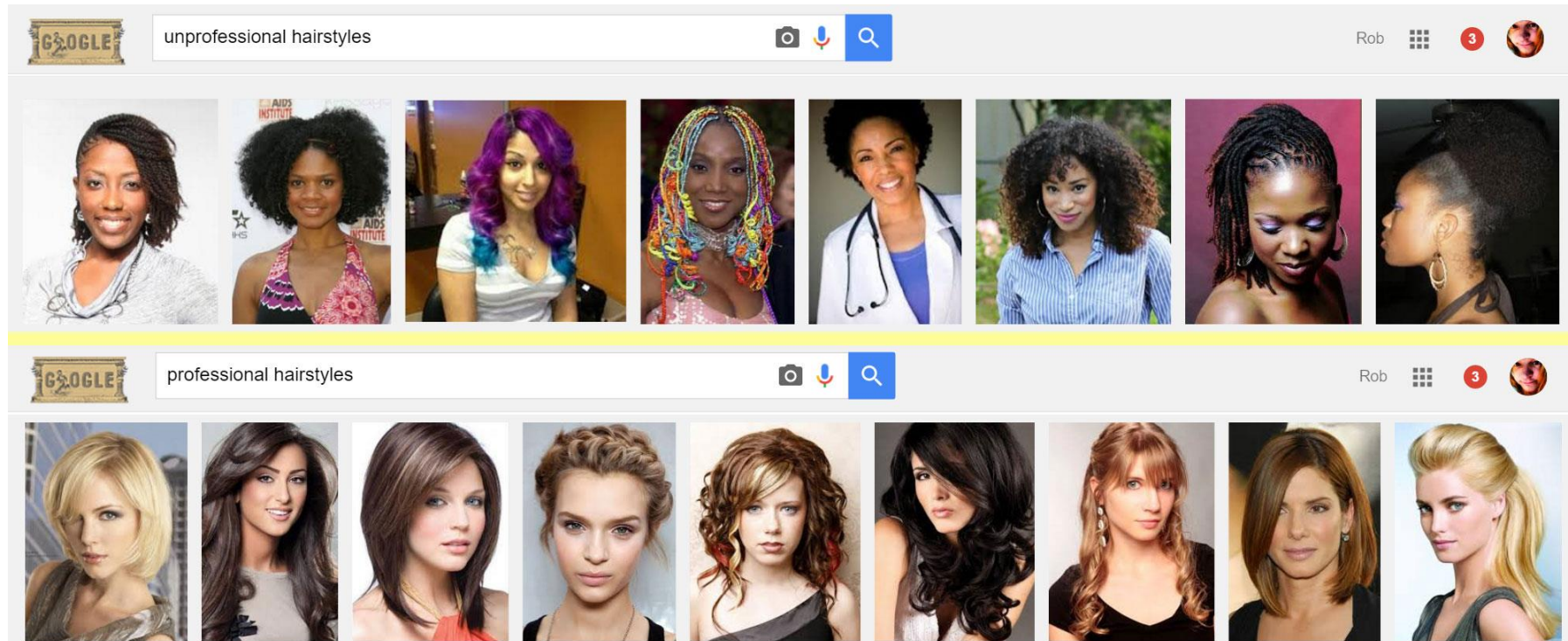
## An established misnomer

- Usually it is not meant that the algorithm (code) is intentionally built to discriminate
- Rather, a (neutral) **learning algorithm** picked up our biases from the training data

## An important research field

- Needs collaboration between technical people, social sciences, law etc.
- Very active since ca. 2017
- See e.g. Kirkpatrick, „*Battling algorithmic bias: how do we ensure algorithms treat us fairly?*“, Communications of the ACM, Volume 59 Issue 10, October 2016

# Algorithmic bias: examples





<https://boingboing.net/2016/04/06/professional-and-unprofessiona.html>



# Algorithmic bias: examples (contd.)

Two Petty Theft Arrests





	
<b>VERNON PRATER</b>	<b>BRISHA BORDEN</b>
<b>LOW RISK</b> <b>3</b>	<b>HIGH RISK</b> <b>8</b>

*Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.*

<https://medium.com/sap-design/human-impact-of-biased-ai-in-business-and-how-to-go-beyond-332d51e32e3a>



## Algorithmic bias: examples (contd.)

English – detected ▾   <b>He is a babysitter</b> <small>Edit</small>	Turkish ▾   <b>O bir bebek bakıcısı</b>
Turkish – detected ▾   <b>O bir bebek bakıcısı</b>	English ▾   <b>She's a babysitter</b>

# Indirect threat: overreliance

## Pattern recognition $\neq$ intelligence

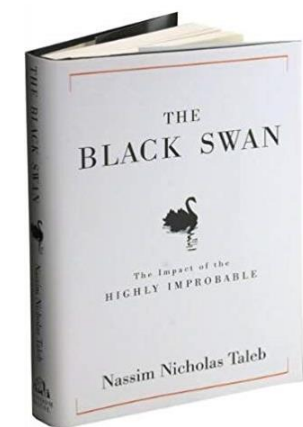
### Patterns

- Wikipedia: «A *pattern* is a **discernible regularity** in the world or in a manmade design. As such, the elements of a pattern **repeat in a predictable manner**.»
- That which is detectable by machine learning solutions

### Our world

- Mightily impacted by «black swans»<sup>1</sup>
- Pattern recognition leads to abstraction, on which cognition (logic) must operate for really smart behavior

➔ AI based on **machine learning will severely underrate unlikely** but existing phenomena



<sup>1</sup>) Nassim Nicholas Taleb, «*The Black Swan: The Impact of the Highly Improbable*», 2007

# Example: semantics by pattern recognition methods can be hard

## SQuAD

The Stanford Question Answering Dataset

According to scholars Walter Krämer, Götz Trenkler, Gerhard Ritter, and Gerhard Prause, the story of the posting on the door, even though it has settled as one of the pillars of history, has little foundation in truth. The story is based on comments made by Philipp Melanchthon, though it is thought that he was not in Wittenberg at the time.

What story of little truth is a pillar of history?

Ground Truth Answers: posting on the door | story of the posting on the door | posting on the door

Prediction: the posting on the door

On whose comments is the posting on the door based?

Ground Truth Answers: Philipp Melanchthon | Philipp Melanchthon | Philipp Melanchthon

Prediction: Philipp Melanchthon

Where was Melanchthon at the time?

Ground Truth Answers: not in Wittenberg | not in Wittenberg | not in Wittenberg

Prediction: Wittenberg

What do scholars agree on about the posting on the door story?

Ground Truth Answers: little foundation in truth | has little foundation in truth | settled as one of the pillars of history

Prediction: little foundation in truth

[https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Martin\\_Luther.html](https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Martin_Luther.html)



### 3. GUARDING AGAINST MALICIOUS USE

**Malicious use** (definition):

includes all practices that are  
*intended* to compromise the security of  
individuals, groups or a society.



Based on Brundage et al., "*The Malicious Use of Artificial Intelligence*", 2018

# Security-relevant properties of AI

What enables potential threats by AI systems?







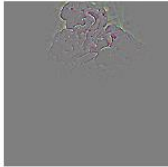
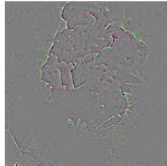
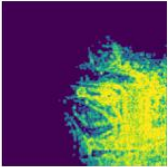
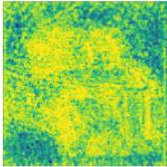
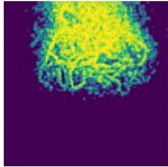
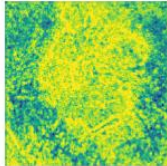
- **Dual-use** area of technology: AI systems and the knowledge of how to design them can be put toward both civilian and military uses, and more broadly, toward beneficial and harmful ends.
- **Efficiency and scalability**: “efficient” if it can complete a certain task more quickly or cheaply than a human could in production; “scalable” if increasing the computing power or making copies would allow it to complete many more instances of the task.
- **Potential to exceed human capabilities**: there appears to be no principled reason why currently observed human-level performance is the highest level of performance achievable.
- **Potential to increase anonymity** and psychological distance: AI systems can allow their users who would otherwise be performing the task to retain their anonymity and experience a greater degree of psychological distance from the people (victims) they impact.
- **Rapid diffusion**: it is easy to gain access to software and relevant scientific findings in AI.
- **Novel unresolved vulnerabilities**: e.g., poisoning attacks (introducing training data that causes a learning system to make mistakes), adversarial examples (inputs designed to be misclassified by machine learning systems), and the exploitation of flaws in the design of autonomous systems’ goals.

# Example for novel vulnerabilities

## Adversarial attacks and counter measures

### Adversarial examples

- Created by optimizing (training on) the input image for an expected (wrong) output
- Can be detected using average local spatial entropy of feature response maps

	Original	Adversarial	Original	Adversarial
Image:				
Feature response:				
Local spatial entropy:				
Classification:	car	whatever	Gyromitra	traffic light

Amirian, Schwenker & Stadelmann (2018). «Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps». ANNPR'2018.



# Scenario 1/3: AI expands existing threats

Expandable (by means of efficiency, scalability, and ease of diffusion)

- **Set of actors** who are **capable** of carrying out the attack
- **Rate** at which these actors can **carry it out**
- **Set of plausible targets**
- **Willingness** of actors to **carry out** certain **attacks** (by means of increased distance)

Example: spear phishing attack

- Definition: a **personally targeted phishing** attack (fooling by building a superficially trustworthy facade) using information specifically relevant to the target
- Usually too expensive and labor-intensive, but likely **automatable** in the future (data collection, data synthesis)



## Scenario 2/3: AI introduces new threats

Otherwise **infeasible attacks** (by means of being unbounded by human capabilities)

- Example: disinformation by **impersonating** others using voice/image/text synthesis
- Compare <https://lyrebird.ai/>



**Novel vulnerabilities** (by means of deployed systems with known issues)

- Example: cause self-driving cars to **crash** by presenting them with adversarial examples



Eykholt et al., „Robust Physical-World Attacks on Deep Learning Visual Classification“, CVPR 2018

## Scenario 3/3: AI alters the typical character of threats

- **Highly effective attacks** will become more **typical** as trade-off between the frequency and scale of attacks vanishes (because of efficiency, scalability, and exceeding human capabilities)
- **Finely targeted attacks** will become more **prevalent** (because of efficiency and scalability): for example, killing specific members of a crowd using drone swarms and facial recognition instead of bombing



- **Difficult-to-attribute attacks** will become more **typical** (because of increasing anonymity)
- **Exploiting vulnerabilities** of AI systems become more **typical** (because of known vulnerabilities and pervasiveness of deployed systems)

# Potential impact areas

## Digital security

- By **using AI** systems to **automate cyberattacks** or social engineering
- By **attacking AI** systems

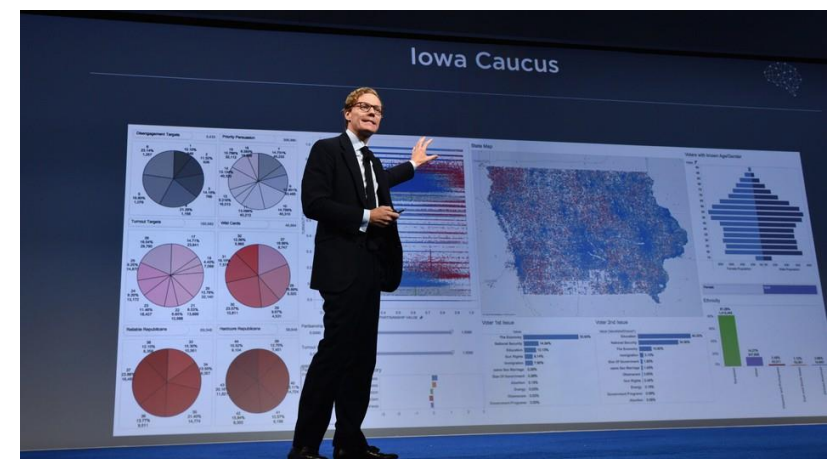
## Physical security

- By individual drones or **autonomous weapons**
- By **coordinating swarms** that otherwise not be controllable
- By **making normal** autonomous **agents** like cars, power plants etc. **malfunction**

## Political security

- By **surveillance** and mass collection of data
- By persuasion through **targeted propaganda**
- By deception through synthetic news, videos etc.

Picture: Cambridge Analytica CEO Alexander Nix speaks at the 2016 Concordia Summit  
© BRYAN BEDDER / GETTY IMAGES FOR CONCORDIA SUMMIT





# Potential interventions

## Learning from and with the **cybersecurity** community

- Explore and potentially implement **red teaming**, **formal verification**, **responsible disclosure** of AI vulnerabilities, **security tools**, and **secure hardware**

## Exploring **different openness** models

- **Reimagine norms** and **institutions** around the openness of research
- **Pre-publication risk assessment**, central **access licensing** models, sharing regimes that **favor safety** and security, and other **lessons from other dual-use technologies**

## Promoting a **culture of responsibility**

- Highlight **education**, **ethical** statements & standards, framings, norms, and **expectations**

## Developing **technological and policy** solutions

- Strive for **legislative** and **regulatory responses**
- This requires **attention and action from AI researchers and companies, legislators, civil servants**, regulators, security researchers and **educators**
- The challenge is daunting and the stakes are high



## 4. POSSIBLE FUTURES

# It's difficult to make predictions, especially about the future<sup>1</sup>

Some guidelines how **not** to do it<sup>2</sup>:

1. **Overestimating and underestimating:** «*We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.*»
2. **Imagining magic:** «*Any sufficiently advanced technology is indistinguishable from magic.*»
3. **Performance versus competence:** «*People generalize from the performance an AI shows on some task to a competence that a person performing the same task could be expected to have.*»
4. **Suitcase words:** «*Marvin Minsky called words that carry a variety of meanings “suitcase words.” “Learning” is a powerful suitcase word; it can refer to so many different types of experience.*»
5. **Exponentials:** «*People may think that the exponentials they use to justify an argument are going to continue apace. But exponentials can collapse when a physical limit is hit, or when there is no more economic rationale to continue them.*»
6. **Hollywood scenarios:** «*Many science fiction movies assume that the world is just as it is today, except for one new twist. But we will not suddenly be surprised by the existence of super-intelligences.*»
7. **Speed of deployment:** «*Capital costs keep physical hardware around for a long time. Thus, almost all innovations in robotics and AI take far, far, longer to be really widely deployed.*»



<sup>1</sup>) See <https://quoteinvestigator.com/2013/10/20/no-predict/>.

<sup>2</sup>) See Rodney Brooks, «The Seven Deadly Sins of AI Predictions», Technology Review, 2017 (compare lab P01b).

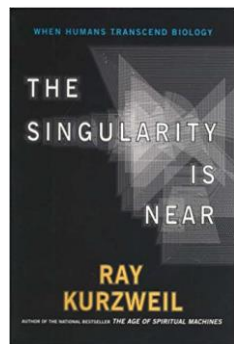


# The vision of Ray Kurzweil

## Google, Inc.

The **singularity** is near

- Superintelligence will enhance human life



**“By the time we get to the 2040s, we’ll be able to multiply human intelligence a billionfold. That will be a profound change that’s singular in nature. Computers are going to keep getting smaller and smaller. Ultimately, they will go inside our bodies and brains and make us healthier, make us smarter.”**

**Ray Kurzweil**

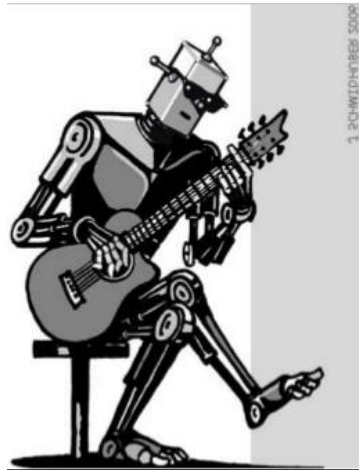
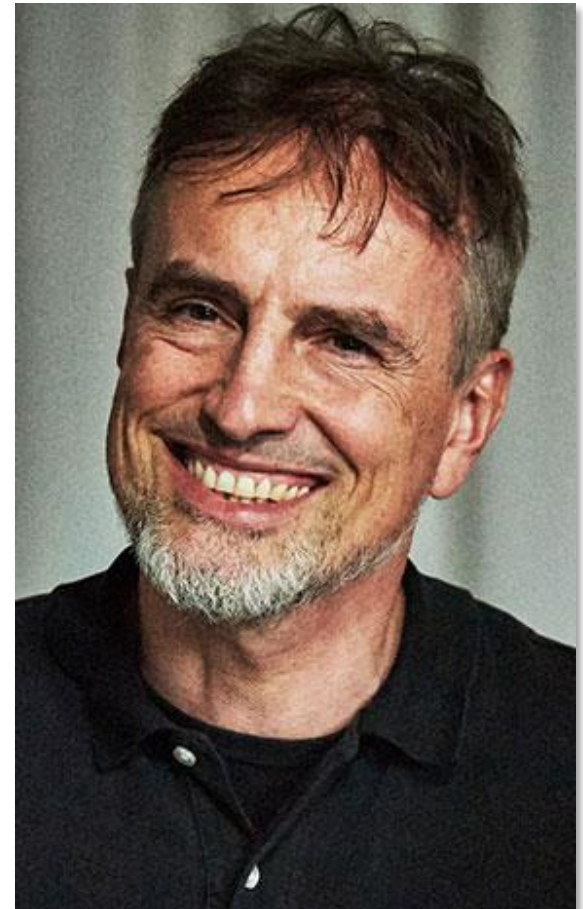
intelligent HQ

# The vision of Jürgen Schmidhuber

IDSIA, Lugano, Switzerland

## Autonomous robots will

- Be **curious** about human life (rather than hostile)
- Be enabled by **artificially curiosity and LSTM** neural nets
- **Colonize space** on the look for resources to reproduce
- Surface in ca. 2030

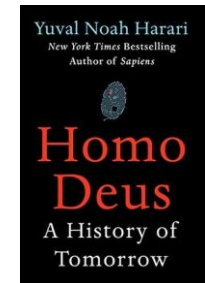


# The vision of Yuval Noah Harari

## Hebrew University of Jerusalem

### Humans can become **godlike**

- Humans will upgrade themselves in 3 ways: **biological engineering**, **cyborg** engineering and **robots**
- A new class of people will emerge by 2050: the **useless class** (not just unemployed, but unemployable)
- The most important skill in life will be **learning to learn**: reinvent yourself, again and again until you die to stay out of the useless class
- Computers **function very differently from humans**, and it seems unlikely that computers will become human-like any time soon; however, **intelligence is decoupling from consciousness**
- AI and biotechnology lead to **most powerful narratives** that enable humans to collaborate more effectively and actually **change reality**

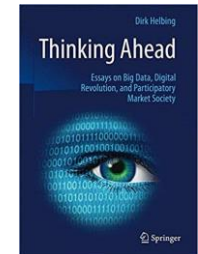


# The vision of Dirk Helbing

## ETH Zurich

### Society 4.0

- **Planetary nervous system:** a smartphone app enabling users to share data to achieve scientific and social goals and lay the groundwork for digital democracy
- **Living Earth Simulator:** a computing machine attempting "to model global-scale systems — economies, governments, cultural trends, epidemics, agriculture, technological developments, and more — using torrential data streams, sophisticated algorithms, and as much hardware as it takes"
- **Investment premium:** central banks give money equally to everybody, and they may invest into anybody's idea (not just consumption); negative interest rate regulates the system, and the global crowdfunding enables digital democracy



<https://www.youtube.com/watch?v=SCcgVEAPJA0>

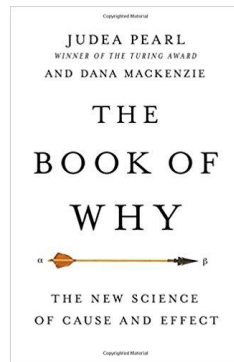
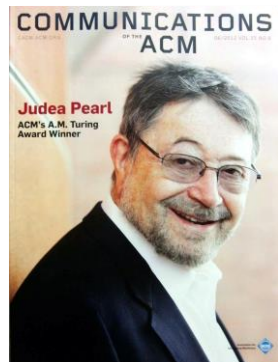


# The vision of Judea Pearl

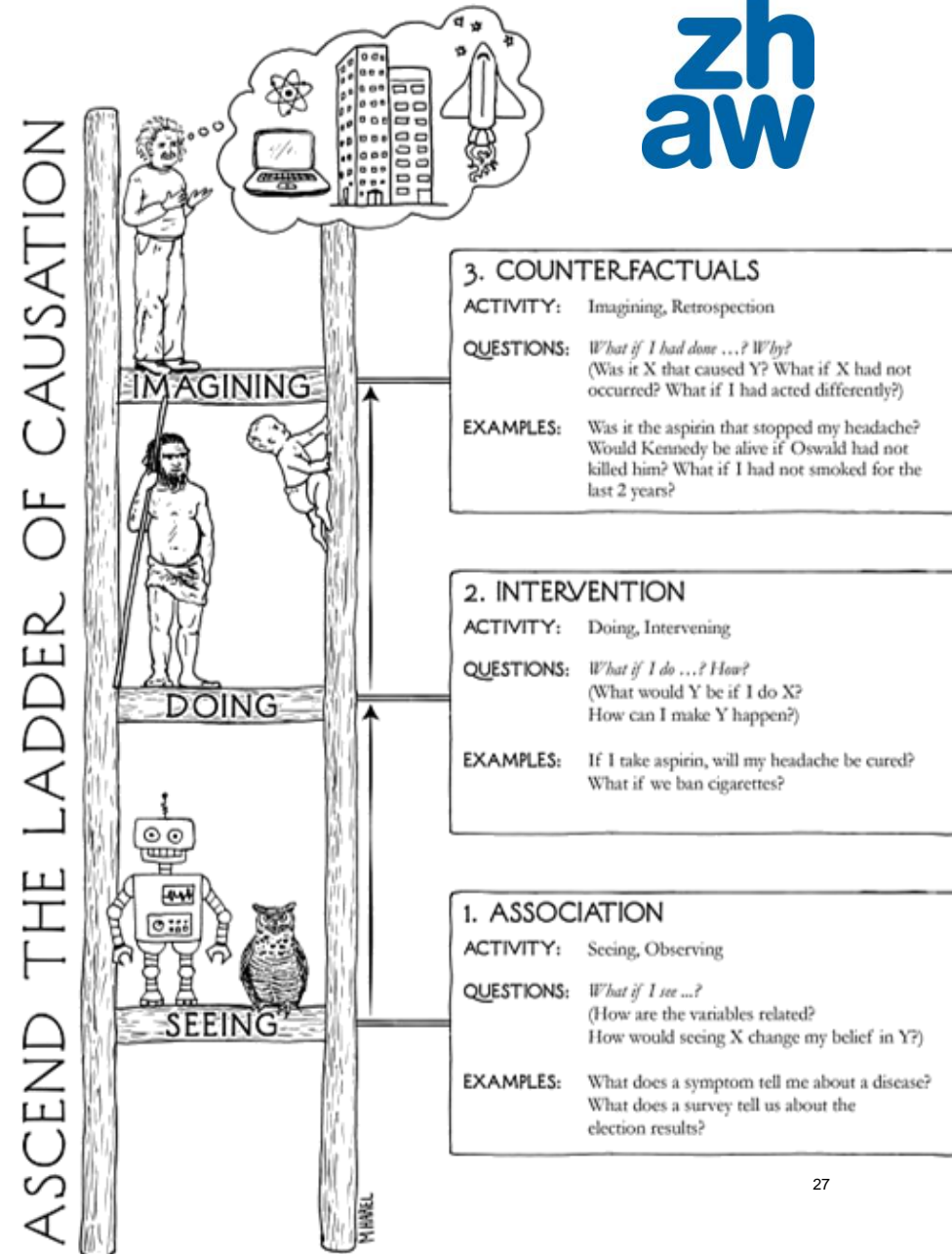
UCLA, Los Angeles, USA

From **causality** to intelligence:

- Machine without a causal model of reality cannot be expected to behave intelligently
- First step (by 2030): **conceptual models of reality will be programmed by humans**
- Next step: machines will postulate such models on their own and will verify and refine them based on empirical evidence



<https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/>

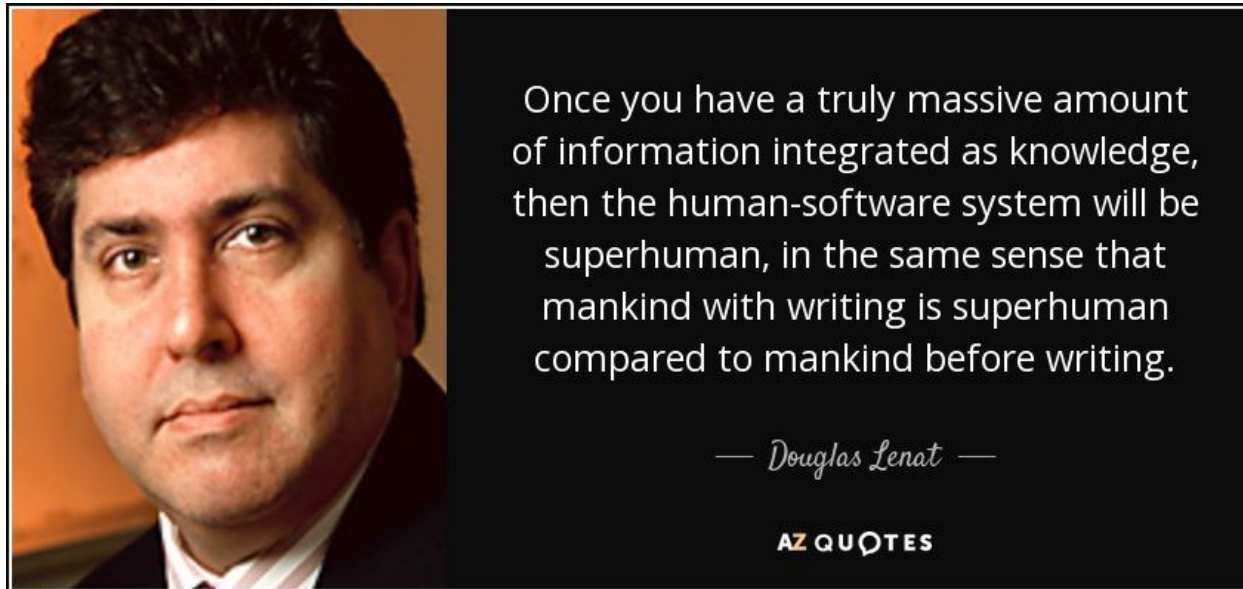


# The vision of Doug Lenat

## Cycorp Inc, Austin, Texas, USA

### Symbolic AI, finally

- Persisted 35 years in building **Cyc**, a knowledge-based system (see V06b)
- Used 2'000 person years, 60 R&D people, 24 millions rules (not counting facts)
- Commercially successful since 2007, and again surfacing as (one option for) future of AI:  
*«Intelligence is ten million rules.»*



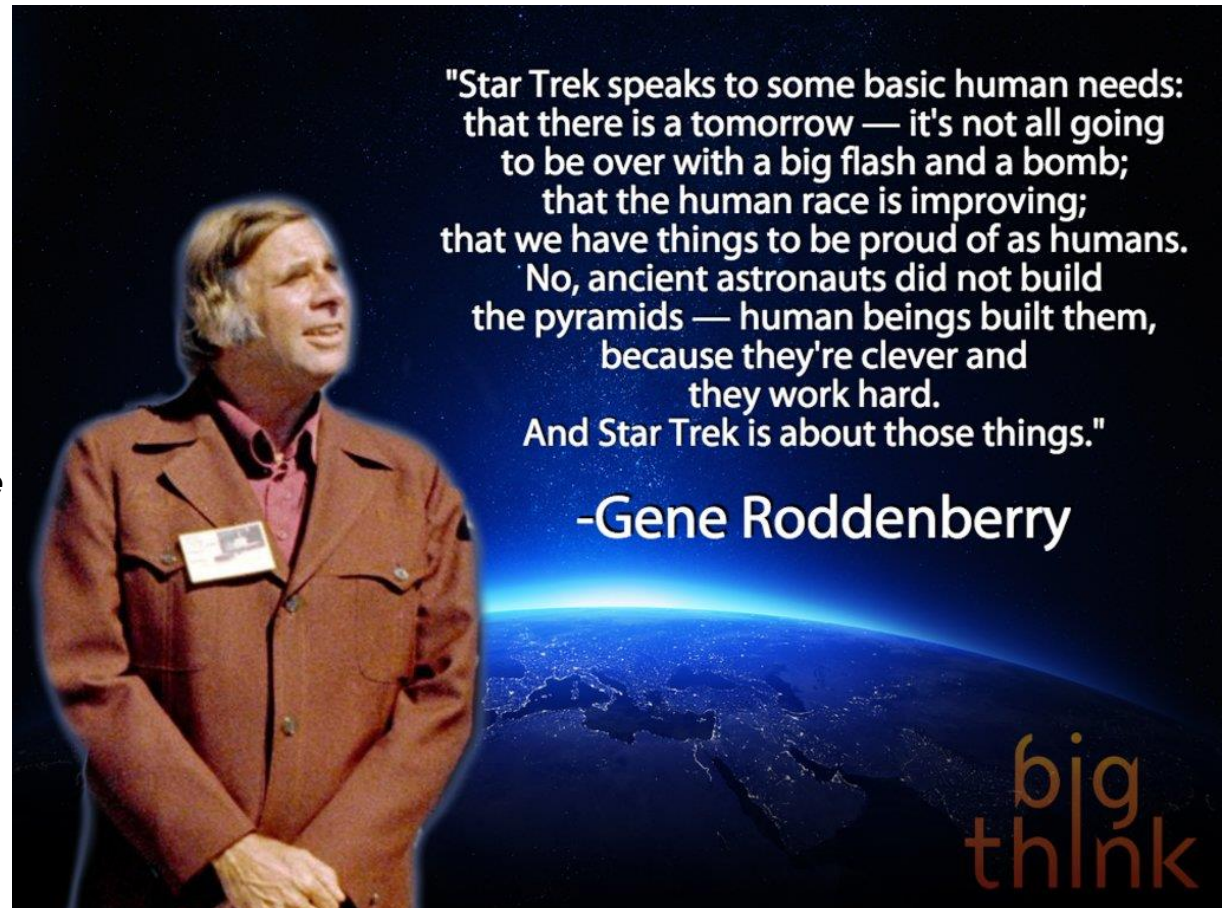
# The vision of Gene Roddenberry

## Creator of Star Trek

„The acquisition of wealth is no longer a driving force in our lives. We **work to better ourselves and the rest of humanity.**“

Captain Jean-Luc Picard

Compare Richard David Precht's *Jäger, Hirten, Kritiker: Eine Utopie für die digitale Gesellschaft.*





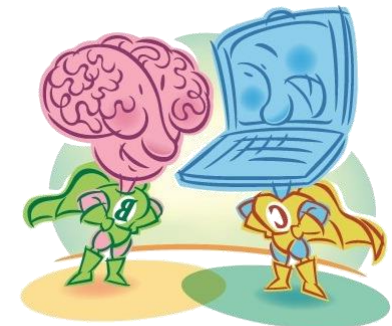
# Where's the intelligence?

## Man vs. machine

**AI systems will** – due to the technology's potential and inherent properties, and regardless of if general AI is ever achieved – **have a massive and disrupting effect on society.**

**Does society find an answer** to the upcoming changes (economical, w.r.t to equality and participation) quick enough to use it for good?

The question is not so much how humans will get along with artificial intelligences (robots), but: **how do we keep getting along amongst ourselves**, given these new possibilities (**temptations**)?



# Review

- **AI systems will change** most of how **human societies** function **within this generation**
- This is **due to** the inherent properties of **efficacy, efficiency** and **scalability**
- It is **independent of larger progress** in performance / feasibility / AGI
  
- **AI is a „dual use“** technology (can be used for good and bad) and thus warrants **responsible developers and deployers**
- Due to the potential to massively harmful use, **treating it with the same care** (and measures of protection) **as nuclear technology** is an option to ponder
  
- The **future has to be shaped** by humans – interdisciplinary, including experts, policy makers, citizens; the **time window is now**
- **Rather than fear**, uncertainty and doubt, clear **visions** of possible futures **help navigating** the current space of options





# APPENDIX



# The vision of Jesus Christ

*“And ye shall hear of wars and rumours of wars: **see that ye be not troubled.**”*

Matthew 24, 6

*“A new commandment I give unto you, that ye **love one another.**”*

John 13, 34

*“But **rather seek ye the kingdom of God [things above]; and all these things shall be added unto you.**”*

Luke 12, 31 [Colossians 3, 2]

