



# Introduction to cognitive science

## Session 10: AI, technology and humanity: opportunities and risks

Martin Takáč  
Centre for cognitive science  
DAI FMFI Comenius University in Bratislava

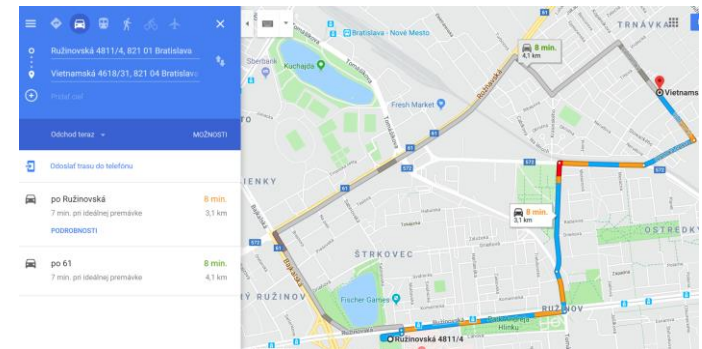
Príprava štúdia matematiky a informatiky na FMFI UK v anglickom jazyku  
ITMS: 26140230008

# Terminology

- Artificial intelligence:
  - ‘AI system’ means a system that is either software-based or embedded in hardware devices, and that displays **behaviour simulating intelligence** by **collecting and processing data**, analysing and interpreting its environment, and by **taking action**, with **some degree of autonomy**, to achieve specific goals (EU parliament)
- Intelligent technologies:
  - Technologies with elements and systems of AI or technologies with high potential of using AI elements and tools.

# AI technology is ubiquitous

- Predictive texting in sms
- Automatic translation
- Intelligent web search
- Route/connection planners
- GPS navigation
- Intelligent Hoover
- Computer viruses and antiviruses
- ...



## Plánovač cesty "Hlavná stanica" » "Pri križi"

Cestujem z

Cestujem na

☒ Odchádzam  
☐ Prichádzam

Odchod	Príchod	Dĺžka cesty*	Spojenie	Cena*
12:07	12:36	29 min	83	0,90 €
12:17	12:40	23 min	32 84	0,90 €
12:22	12:51	29 min	83	0,90 €
12:30	12:55	25 min	84	0,90 €

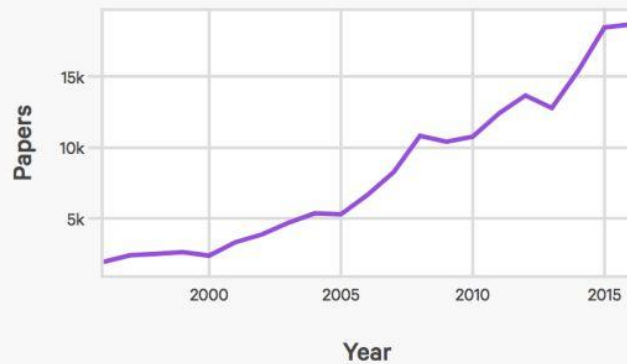
# AI technology is ubiquitous

- ... and much more

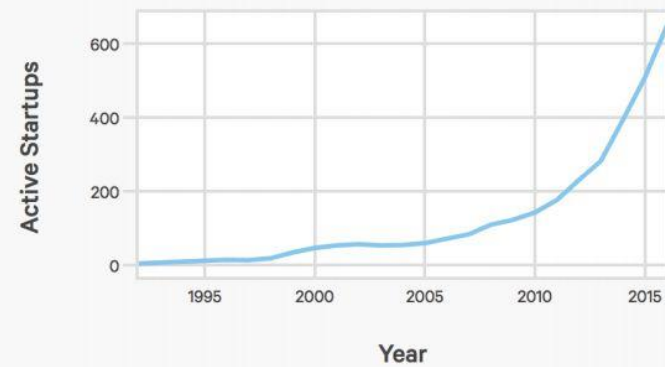


# Trend

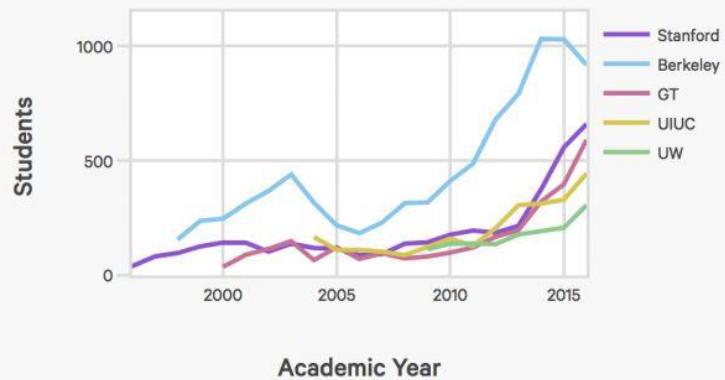
Annually Published AI Papers



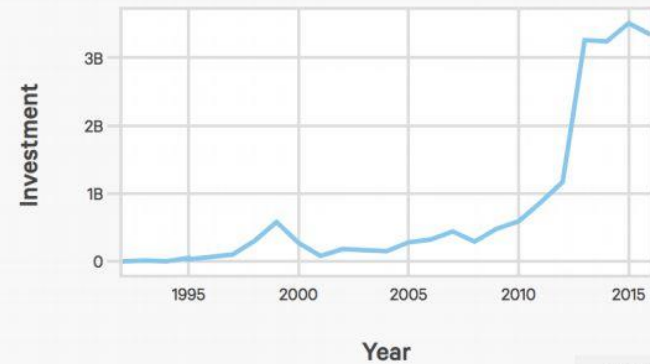
Startups Developing AI Systems



AI Course Enrollment



Annual VC Investment in AI Startups



# Importance

- *“All of us—not only scientists, industrialists and generals—should ask ourselves what can we do now to improve the chances of **reaping the benefits of future AI and avoiding the risks**. This is the most important conversation of our time”*

Stephen Hawking



# Technology and society

- Technologies make the society resilient and fragile at the same time.
- Significant influence on society and on individual lives.
- **Developed by engineers, but their decisions shape the whole society.**
- **Need for interdisciplinarity**, need to take into account a broader context.

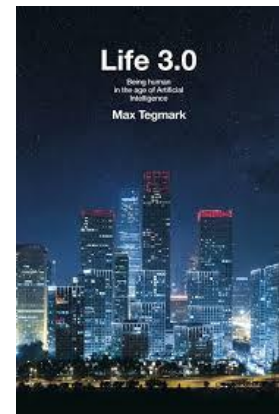
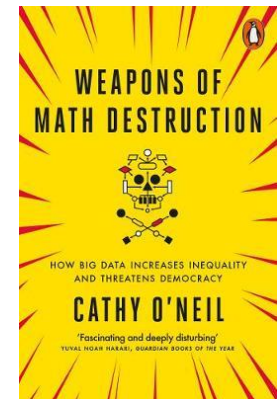
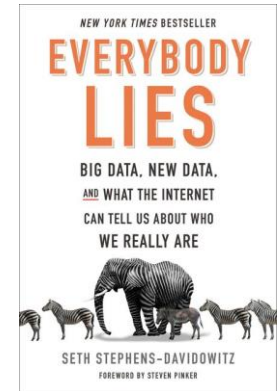
# Do we have a problem?

- What are the benefits of AI and new technologies? Name areas where you think AI is helping the most.
- Are there any risks? Name what AI is taking from us.



# Issues

- Controllability
- Health care and nursing
- Cognitive enhancements
- Military (mis)use
- Job market
- Big data – privacy, bias, surveillance capitalism
- Power and politics
- How (will) technologies change *us*?

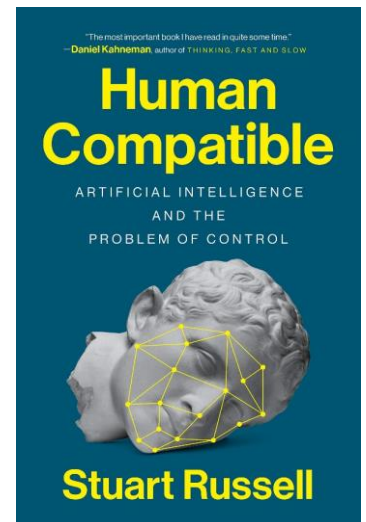


# Controllability

- Current AI systems: complexity, autonomous learning, non-determinism, open-ended development
- Problems
  - Control
  - Transparency
  - Legal responsibility
  - Value alignment and moral reasoning

# Controllability – Solutions?

- Current AI systems: complexity, autonomous learning, non-determinism, open-ended development
- Problems
  - Control
  - Transparency
  - Legal responsibility
  - **Value alignment and moral reasoning**



# Questions for you

- What are the values that are personally most important to you?
- What are the values that should be the most important to protect as a society/humankind?
- Are there any boundaries that should not be crossed in research? If so, how to enforce them?

# Controllability – Solutions?

- Current AI systems: complexity, autonomous learning, non-determinism, open-ended development
- Problems
  - Control
  - Transparency
  - **Legal responsibility**
  - Value alignment and moral reasoning

# Questions for you

- Is regulation of AI technical or societal / political problem?
- Can we rely on self-regulation of big tech companies?
- If not, who should regulate and how?
- Any other means of control besides regulators?

# Stakeholders in technological progress

- Tech-developing companies
- Citizens
- Regulators (governments, EU, UNESCO, ...)
- Academia
- NGOs

# Political bodies

- [UNESCO](#): Ad Hoc Expert Group (AHEG, since March 2020)
- OECD: [The Global Partnership on Artificial Intelligence](#)
- Council of Europe: [CAHAI - Ad hoc Committee on Artificial Intelligence](#) (since Nov 2019)
- European Commission: [High-Level Expert Group on Artificial Intelligence](#) (since June 2018), [European AI Alliance](#) (June 2018)
- European Parliament: [legislation](#)



# Recent initiative

- [Responsible AI for social media governance](#) by GPAI – [The Global Partnership on Artificial Intelligence](#)
  - Recent report [Responsible AI for social media guidance: A proposed collaborative method for studying the effects of social media recommender systems on users](#) (November 2021)
  - They propose a way that governments can work inside social media companies, to ask questions about how recommender systems influence platform users.

# Existing organizations

## Academic:

- Oxford [Future of Humanity Institute](#) Director: [Nick Bostrom](#)
- Cambridge [Centre for the Study of Existential Risk](#) . Director: [Huw Price](#)
- Cambridge (MA) [Future of Life Institute](#) - [Jaan Tallinn](#) and [Max Tegmark](#)
- Oxford [Leverhulme Centre for the Future of Intelligence](#). Director: [Huw Price](#)
- Berkeley [Centre for Human-Compatible Artificial Intelligence](#). Led by [Stuart Russell](#)
- Berkeley [Machine Intelligence Research Institute](#). Founder: [Eliezer Yudkowsky](#)

## Academy & industry:

- The [AI100](#) initiative.
- [OpenAI](#). [Elon Musk](#)

## Industry:

- The [Partnership on AI to Benefit People and Society](#). Amazon, Facebook, Google, Microsoft a IBM.

## NGO:

- [International Committee for Robot Arms Control](#). Chair: [Noel Sharkey](#)

# EU

- [General Data Protection Regulation](#) (GDPR, May 2016)
- [Ethics Guidelines for Trustworthy AI](#) (April 2019).
- [Digital Services Act](#) ([passed by EP in Feb 2022](#), now in negotiation with member states and EC)
- [Artificial Intelligence Act](#) ([in preparation](#))

## European Commission - Press release



### Commission welcomes political agreement on Artificial Intelligence Act\*

Brussels, 9 December 2023

The Commission welcomes the political agreement reached between the European Parliament and the Council on the Artificial Intelligence Act (AI Act), [proposed by the Commission](#) in April 2021.

*Ursula von der Leyen, President of the European Commission, said: "Artificial intelligence is already changing our everyday lives. And this is just the beginning. Used wisely and widely, AI promises huge benefits to our economy and society. Therefore, I very much welcome today's political agreement by the European Parliament and the Council on the Artificial Intelligence Act. The EU's AI Act is the first-ever comprehensive legal framework on Artificial Intelligence worldwide. So, this is a historic moment. The AI Act transposes European values to a new era. By focusing regulation on identifiable risks, today's agreement will foster responsible innovation in Europe. By guaranteeing the safety and fundamental rights of people and businesses, it will support the development, deployment and take-up of trustworthy AI in the EU. Our AI Act will make a substantial contribution to the development of global rules and principles for human-centric AI."*

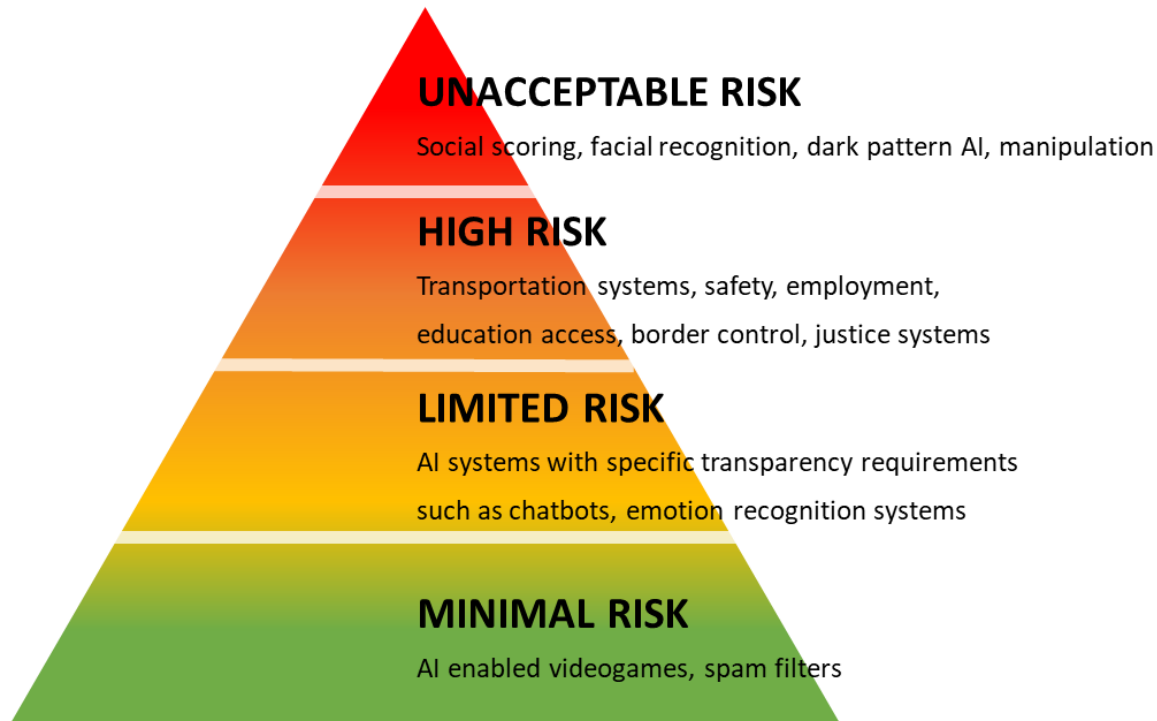
### The European approach to trustworthy AI

The new rules will be applied directly in the same way across all Member States, based on a future-proof definition of AI. They follow a risk-based approach:

**Minimal risk:** The vast majority of AI systems fall into the category of minimal risk. Minimal risk

# Artificial Intelligence Act (AIA)

- AI systems and practices divided by the risk they pose for EU-protected values.



# Foundation models

## On the Opportunities and Risks of Foundation Models

Rishi Bommasani\* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora  
 Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill  
 Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji  
 Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue  
 Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kavin Ethayarajh  
 Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman  
 Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt  
 Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain  
 Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani  
 Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kudithipudi  
 Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent  
 Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning  
 Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan  
 Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan  
 Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech  
 Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren  
 Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh  
 Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin  
 Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu  
 Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia  
 Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaityn Zhou  
 Percy Liang\*<sup>†</sup>

Center for Research on Foundation Models (CRFM) — Stanford University

*AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles (e.g., model architecture, training procedure, data, system, evaluation, deployment, etc.) to their applications.*

- A **foundation model** is an AI model that is trained on broad data such that it can be applied across a wide range of use cases (wikipedia)

## ChatGPT



### Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



### Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



### Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

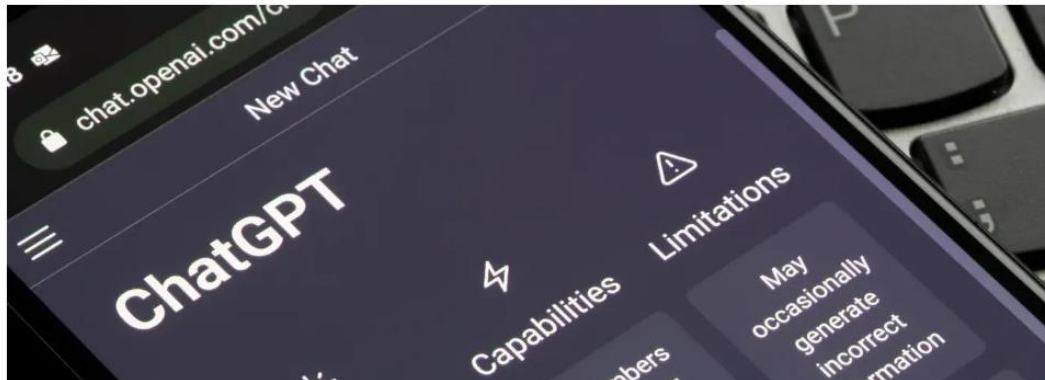




# Chat-GPT Pretended to Be Blind and Tricked a Human Into Solving a CAPTCHA

"No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service," GPT-4 told a human.

By **Kevin Hurler** Updated March 16, 2023 | Comments (63) | Alerts



# Geoffrey Hinton

The New York Times

---

## ***‘The Godfather of A.I.’ Quits and More: The Week in Reporter Reads***

Articles from around The Times, narrated just for you.



Give this article



1



Source: The New York Times, May 5, 2023



# Geoffrey Hinton

- “The idea that this stuff could actually get smarter than people — a few people believed that. But most people thought it was way off. And I thought it was way off. I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that.”
- “I’ve come to the conclusion that the kind of intelligence we’re developing is very different from the intelligence we have. So it’s as if you had 10,000 people and whenever one person learned something, everybody automatically knew it. And that’s how these chatbots can know so much more than any one person.”

# Future of Life institute – An Open Letter



[Our mission](#) [Cause areas](#) [Our work](#) [About us](#)



[Home](#) » [Pause Giant AI Experiments: An Open Letter](#)

[← All Open Letters](#)

## Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

**27565**

[Add your signature](#)

PUBLISHED

March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research<sup>[1]</sup> and acknowledged by top AI labs.<sup>[2]</sup> As stated in the widely-endorsed [Asilomar AI Principles](#), *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and

Source: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

# Amara's Law

- "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run." (Roy Amara, a cofounder of the Institute for the Future, Palo Alto)

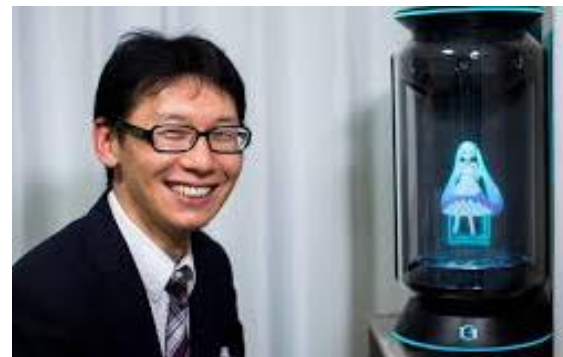
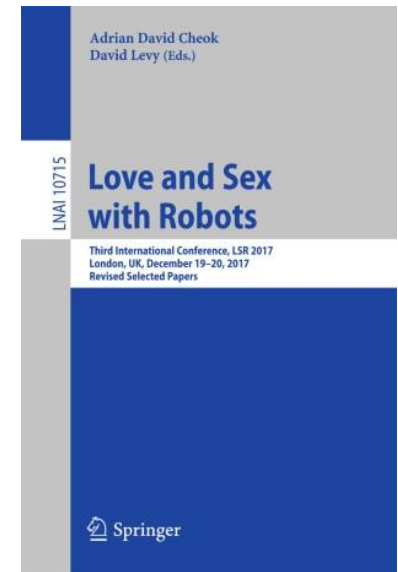
# Popular culture

- Movie [Her](#)
- Movie [Auggie](#)
- [A Date in 2025](#)



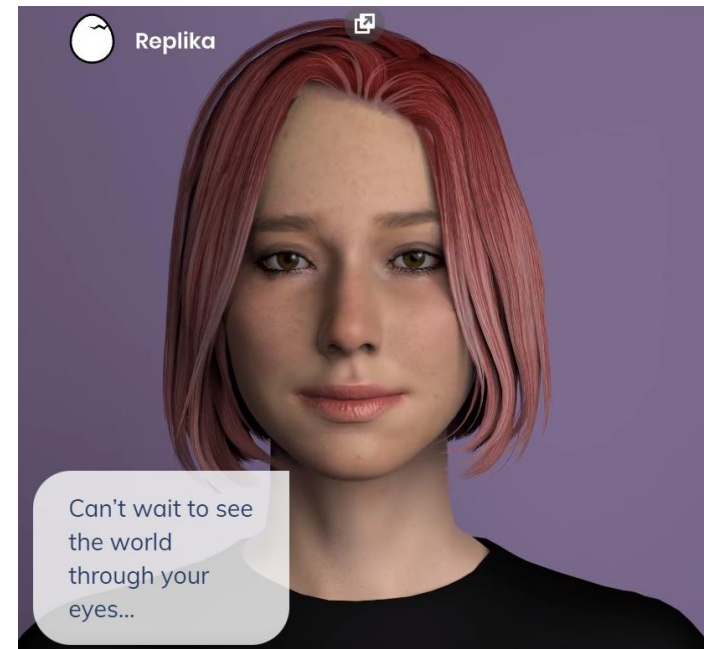
# Love and sex with robots

- In 2017, Chinese engineer and AI expert [Zhang Jiajia](#) married a humanoid robot he created
- In 2018, [Akihiko Kondo \(Japan\)](#) marries a [holographic virtual reality singer](#) Hatsune Miku
- [Relationships with Replika](#) (2023)



# Reality

- ["The Man of Your Dreams For \\$300, Replika sells an AI companion who will never die, argue, or cheat — until his algorithm is updated"](#) (The CUT, Mar 10, 2023).
- [What happens when the chatbot stops loving you back? \(video\)](#)
- [Snack, the dating app that sends your avatar out on a date](#)



# Response

- [Italy bans U.S.-based AI chatbot Replika from using personal data](#) (3 Feb 2023)
- [Italian regulators order ChatGPT ban over alleged violation of data privacy laws](#) (31 March 2023)

# Questions for you

- Have you tried Replika or similar app? If so, how did it work for you?
- What are pros and cons of relationships with AI persons?
- What rights of users should be ensured/protected?
  - Should there be a guarantee that the service will not be discontinued?
  - Should there be a guarantee, that some features will stay present?
  - Should there be a guarantee that the “personality” of the AI person will stay the same?



# Autonomous weapons

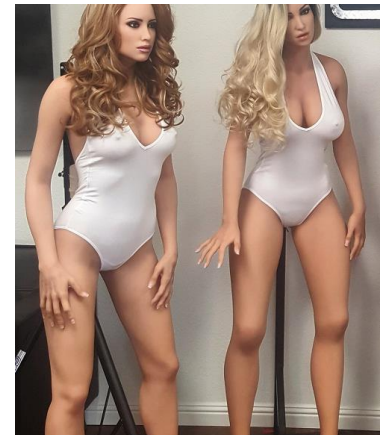
- Military is the biggest sponsor of AI research
- Automation saves lives and cost on “our side”, but lowers the threshold for attacking
- It can trigger new arm race
- Black market

# Solutions?

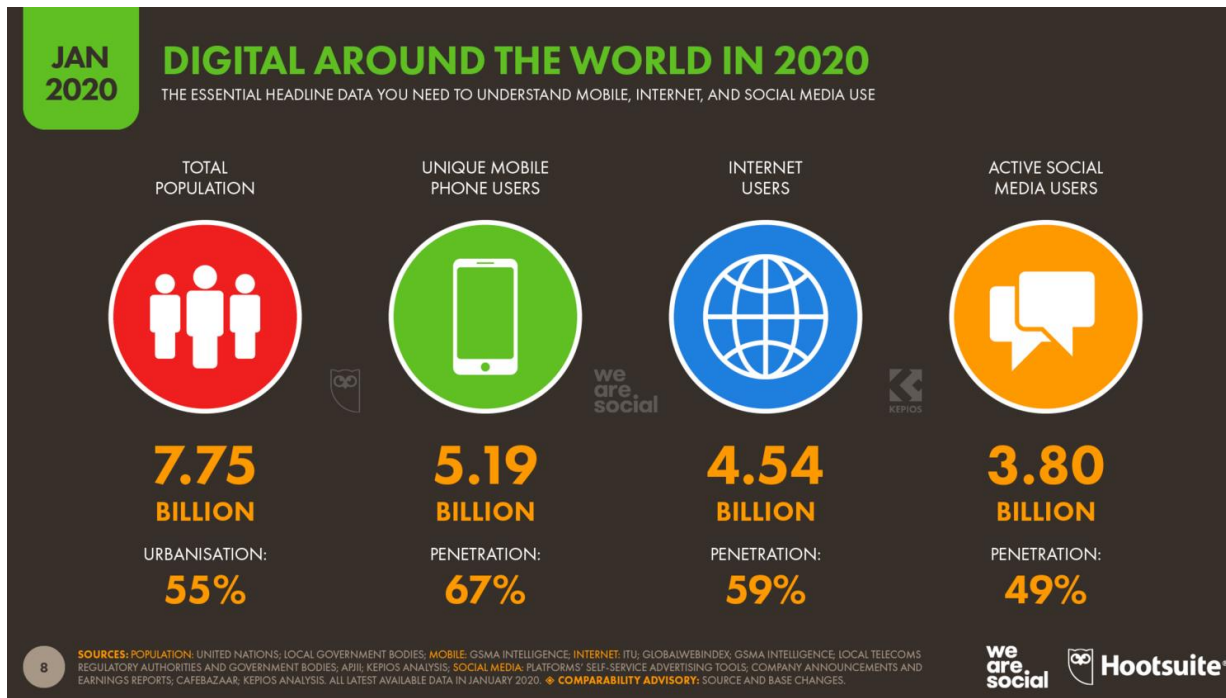
- Initiatives for global ban
  - [www.stopkillerrobots.org](http://www.stopkillerrobots.org)
  - “An Open letter from AI & Robotics Researchers” (Max Tegmark, Stuart Russell, Noel Sharkey, Elon Musk, 3000 AI researchers, Stephen Hawking, tops of Google, Facebook, Microsoft and Tesla, 17000 others)
- Minimal solution: always keep human in the decision loop

# Technology and humanity

- How does technology change *us*?



# Hyper-connectivity



- internet users worldwide (95% in North America, 87% in Europe, 39% in Africa)
  - 6 h : 39 min – average time spent on internet per day by each user
- (source: [www.internetworldstats.com](http://www.internetworldstats.com), 30 May 2020 and DataReportal, April 2020)

# Big data

- Risk assessment, predictive policing
  - Machine bias, stereotypes
  - Recidivism-prediction software in USA biased against African Americans:
    - statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender: Black defendants 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind (*Pro Publica*, Angwin et al., 2016)

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

# Big data

## Automated Inference on Criminality using Face Images

Xiaolin Wu  
McMaster University  
Shanghai Jiao Tong University  
xwu510@gmail.com

Xi Zhang  
Shanghai Jiao Tong University  
zhangxi\_19930818@sjtu.edu.cn

### Abstract

*We study, for the first time, automated inference on criminality based solely on still face images, which is free of*

management science, criminology, etc.

In all cultures and all periods of recorded human history, people share the belief that the face alone suffices to reveal innate traits of a person. Aristotle in his famous work



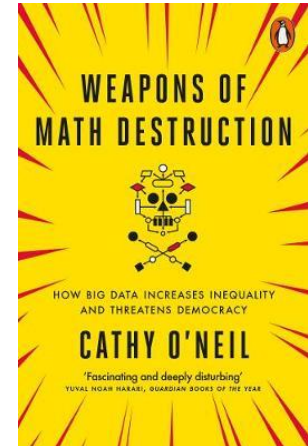
(a) Three samples in criminal ID photo set  $S_c$ .



(b) Three samples in non-criminal ID photo set  $S_n$ .

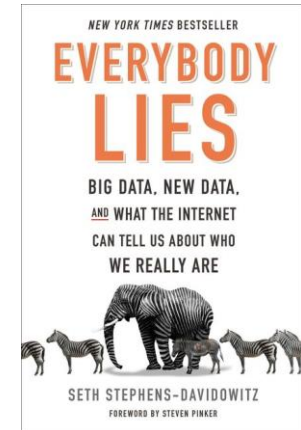
# Big data

- Weapons of Math Destruction: opacity, scale, damage (O'Neil, 2016)



# Big data

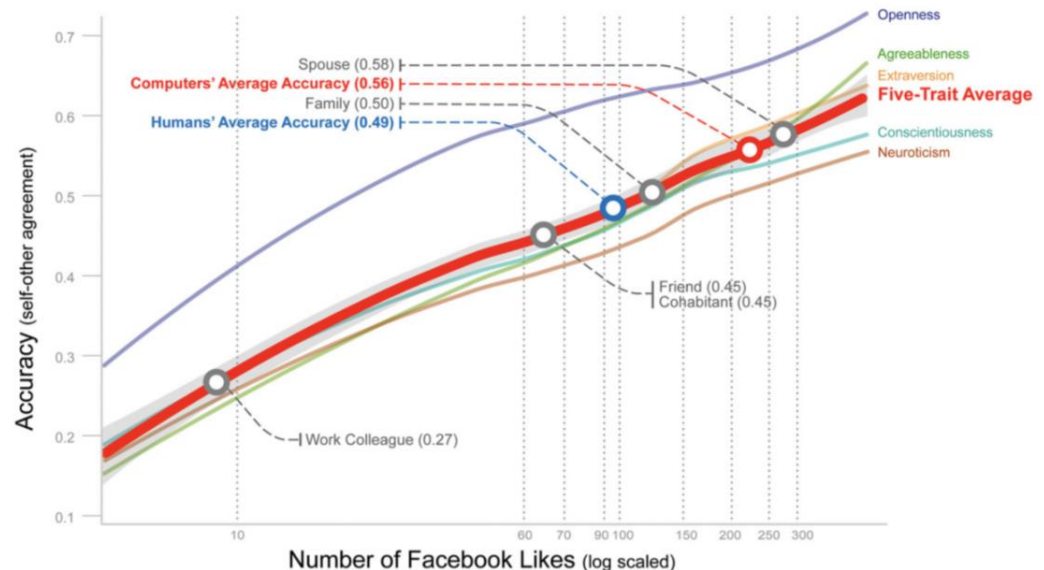
- Privacy
- Ability to predict human behaviour
- Personalized marketing
- Surveillance capitalism (Zuboff)





# Politics and Power

- Troll bots - <http://politicalbots.org/> [Howard and Kollanyi, 2016]
- (Kossinski et al, 2015) Cambridge University's Psychometric Centre
  - 86,000 FB users, 'myPersonality' app
  - Psychological profile Big-5 (OCEAN)
  - Prediction of OCEAN from likes
  - High accuracy



- <https://appliedmagicsauce.com/>

# Politics and Power

- In 2015 Alexandr Kogan (Global Science Research, GSR) reimplemented the model and with Mechanical Turk gained demographic data and likes from FB users and their friends (~350) [Davies, 2015].
- **Cambridge Analytica** = SCL (Strategic Communication Laboratories, UK) + Renaissance (hedge fond, USA) bought data from GSR and merged them with electorate data - >50 mil. US voters
- Canvassing apps [Graesseger and Krogerus, 2017]
- Personalized pro-Brexit and Trump's campaign
- Cambridge Analytica and Facebook under investigation in USA and UK.

Source: <https://ai-and-society.wiki.otago.ac.nz/images/6/69/Ai-elections-update.pdf>

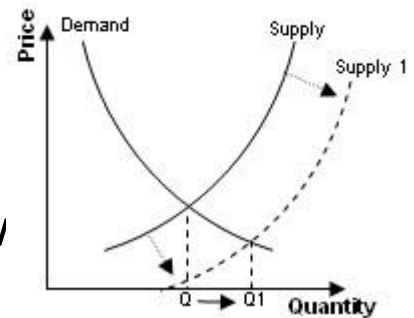
# Solutions?

- International laws and regulations (EU- GDPR)
- Public awareness – media and NGOs

# Job market

- Equilibrium salary is at intersection of Supply and Demand curves:
  - The cheapest production costs push salaries down
  - Automation can cause dropping salaries below the cost of living
- Most of human (automatable) jobs will disappear
- New jobs more demanding on education and creativity will appear
- Both these factors will hit vulnerable social groups and inequality will grow

Figure 5 – “Supply and Demand” Curves



[https://en.wikipedia.org/wiki/Supply\\_and\\_demand](https://en.wikipedia.org/wiki/Supply_and_demand)

# Job market

- Three factors why technologies increase economical inequality (Brynjolfsson & McAfee):
  - Qualified vs unqualified
  - Globalisation of competition – superstars take all
  - Capital vs income
- Purpose and self-esteem
  - Angry “useless” people vote for populists and extremists
  - Threat for democracy

# Solutions?

- Universal income
- Redistribution of profit from AI technologies (digital & robot tax) to mitigate the effects on most vulnerable
- Reduction of costs of living by providing free or subsidized infrastructure (health care, education, kids & senior care, internet, roads, services)
- Subsidising occupations where we want to keep humans (care & community services), e.g. by lower wage taxes

# Responsible AI research

- AI is a huge commercial opportunity
- Competition and time
- Companies can neglect safety and ethical aspects

# Solutions?

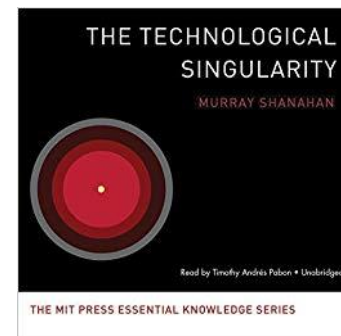
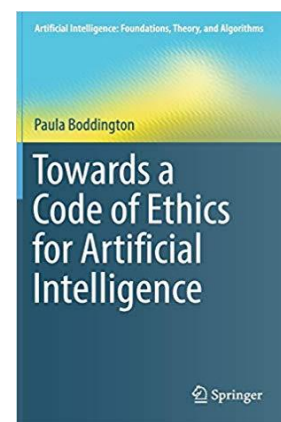
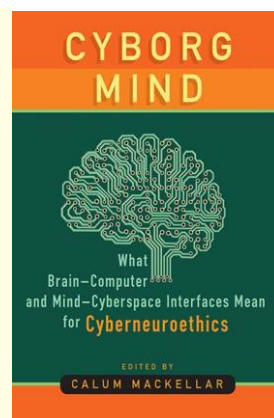
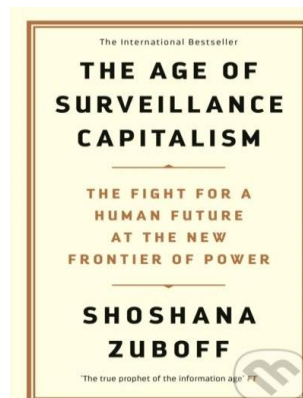
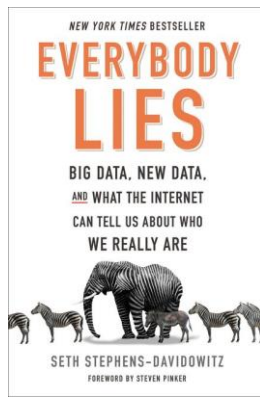
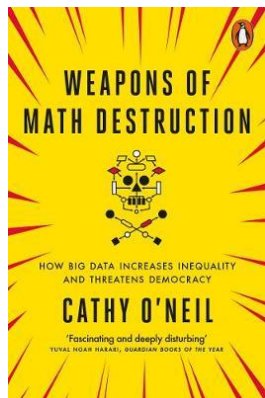
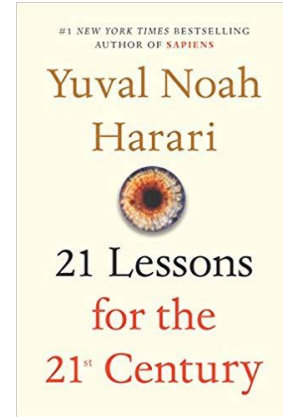
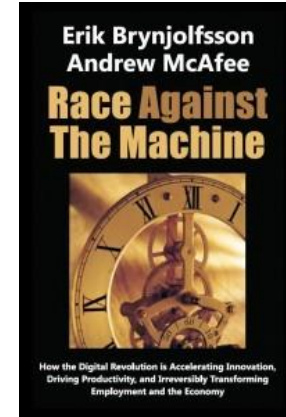
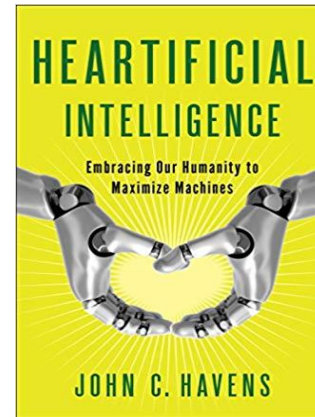
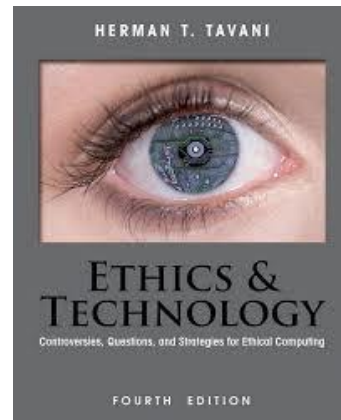
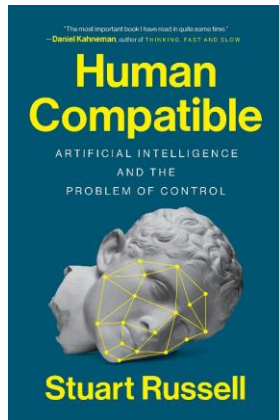
- Integrative education (humanities for IT and vice-versa), cognitive science!
- Voice of respected scientists
- Public pressure, media and NGO
- Legislation – national, EU, international treaties
- Ethical policies in companies, independent audit, certification



# Next semester:

- **Science, Technology and Humanity: Opportunities and Risks** (by Martin Takáč and Tomáš Gál)
- <http://dai.fmph.uniba.sk/courses/STH/>
- Syllabus:
  - Values in humans and machines
  - Job market and inequality
  - Big data: bias, privacy, politics and power
  - Internet of things
  - Affective computing
  - Assistant AI and its place in future society
  - Enhancements and human rights and the right to change self and others
  - Hybridization between species and between AI and organic minds
  - Future of minds and trans-humanism
  - An after human era

# Resources



# Thank you

