Overview: Ethics of AI

Vincent C. Müller

TU Eindhoven

www.sophia.de – 12.2021

Outline

1. Background:
   AI Ethics structure

2. Current debates

3. Policy of AI

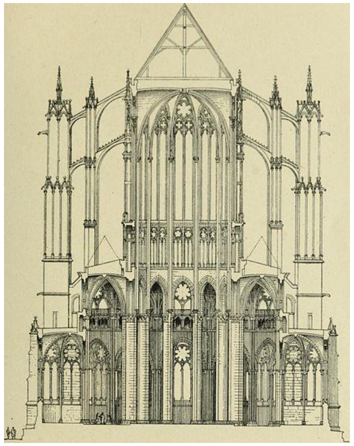
4. What is cold & what is hot

1) Structure: “Ethics of AI & Robotics”

(Stanford Encyclopedia of Philosophy - https://plato.stanford.edu/entries/ethics-ai/)

1. Introduction

2. Main Debates



2.1-2 Data: Privacy & Manipulation

2.3-4 Epistemology: Opacity & Bias

2.5-7 Robot Ethics: Automation, Interaction, Autonomy

2.8-9 Concepts (Agency, Responsibility, Autonomy …)

2.10 Singularity (Superintelligence)

2) Current debates in AI ethics [examples]

- **Criteria**
  - Theoretical interest (‘ethical problem’)
  - Practical interest (utility, injustice)

2.1 Data, all data, all personal data?

- Data
  - Our life is largely digital
  - All classic data is now digital (banking, government, medical, …)
  - More sensors: ‘smart cities’, ‘smart homes’, ‘smart phones’, ‘wearables’, ‘quantified self’, ‘Internet of Things’, …
  - Classic surveillance
- Advanced data analysis, Big Data, personal identification, prediction, …

2.1-2 Surveillance & Manipulation of Behaviour

- The data train we leave behind is how our ‘free’ services are paid for - “surveillance is the business model of the Internet” (Schneier 2015); “surveillance capitalism” (Zuboff 2019).
- The attention economy (Google, Facebook, big 5) is based on deception, exploiting human weaknesses, generating addiction, and manipulation (Harris 2016, J. Williams 2018)
- Manipulation of action beyond economic aims
- Manipulation of text, images, video, …

2.3 Opacity (shallow)

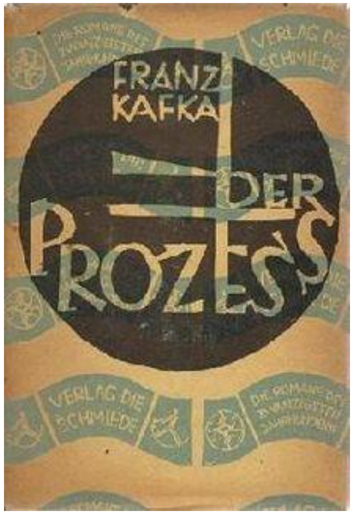
- Governance decisions are made by the automated big-data system (machine-learning, AI)
- Human users in-the-loop / on-the-loop / out-of-the-loop
  - Subjects can be constrained, manipulated or ‘nudged’
  - Decisions cannot be challenged
  - Human users are not responsible and subjects not in control

Standard opacity: “Black Box AI”

- Machine Learning Systems
  - Supervised, semi-supervised (e.g. reinforcement), unsupervised → ‘patterns’
- ‘Black Box’
  - We do not know how the machine generated the patterns
    - → ‘fooling problem’
  - We may find patterns we were not looking for – that nobody knew
- Black Box → Black Box Problem
  - Context: Justification of action (judgment)
  - A “process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process” (Humphreys 2009, 618)
- Opacity for the experts
  - → “explainable AI” (ML) (framework: Zednik 2020)

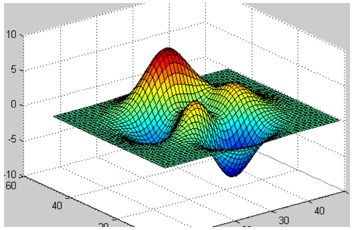
Opacity and politics

- “algorithmic governance creates problems for the moral or political legitimacy of our public decision-making processes” (Danaher 2016)
- “How the Enlightenment Ends”. AI is “a potentially dominating technology in search of a guiding philosophy” (Kissinger 2018)
- AI may leads to a Kafka-style impenetrable suppression system (Cave 2019)



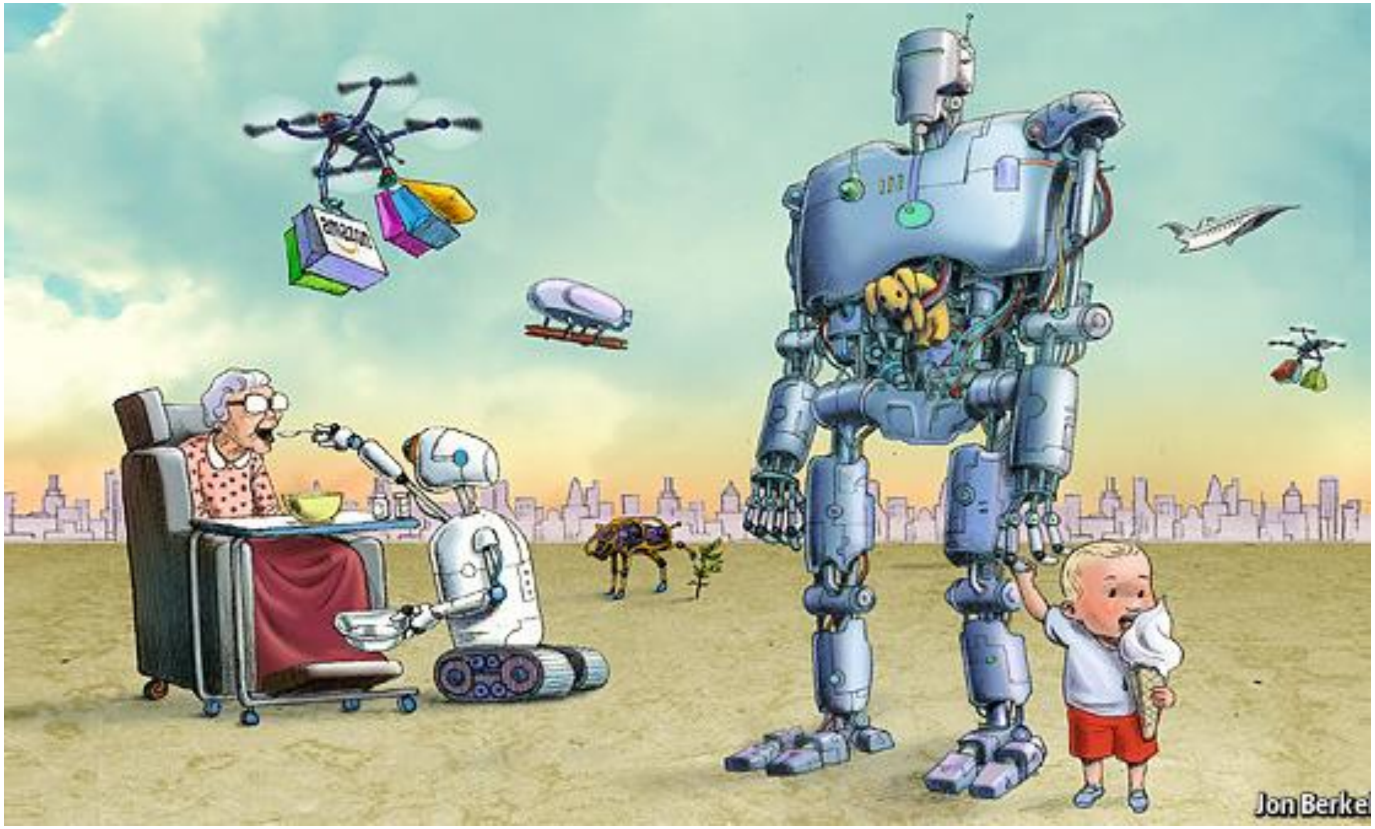
2.4 Bias

- A decision on what is fair implies a decision of what are the relevant characteristics
- Judging by an irrelevant characteristic (e.g. a job candidate by skin colour) is using a bias and discriminatory
- Machine learning learns past bias
- Machine learning is opaque to users and makers
- Bias → unfair?



“Rise of the robots”

(*The Economist* 28.3.2014)

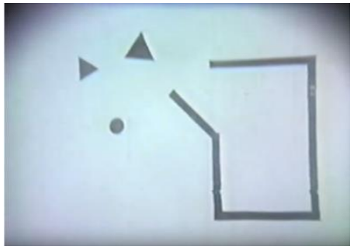


2.5 Interaction with Machines

- Issues: Deception, human dignity & respect for humans (+safety in HRI)
  - … exploiting human weaknesses?
  - A question of power?
- 2.5.2 Example a) Care Robots
- 2.5.3 Example b) Sex Robots

… all too human

- “I find people willing to seriously consider robots not only as pets but as potential friends, confidants, and even romantic partners. We don’t seem to care what their artificial intelligences ‘know’ or ‘understand’ of the human moments we might ‘share’ with them... the performance of connection seems connection enough” (Turkle 2012, 9).
- "While toasters are designed to make toast, social robots are designed to act as our companions.” (Darling 2016, 216)



Heider, F; Simmel, M (1944). "An experimental study of apparent behavior". American Journal of Psychology. 57: 243–259.

2.5.1 Care

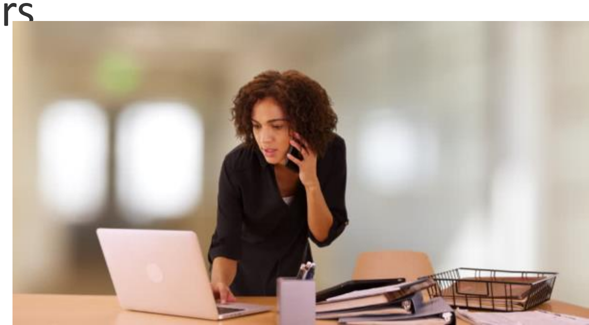
- Robots in health care - de-humanising care?
- Practically: Lifting patients, transporting material, eating with robot arm, robots for comfort
- Is the dystopia ‘care robots’? Automated care? Or non-care?

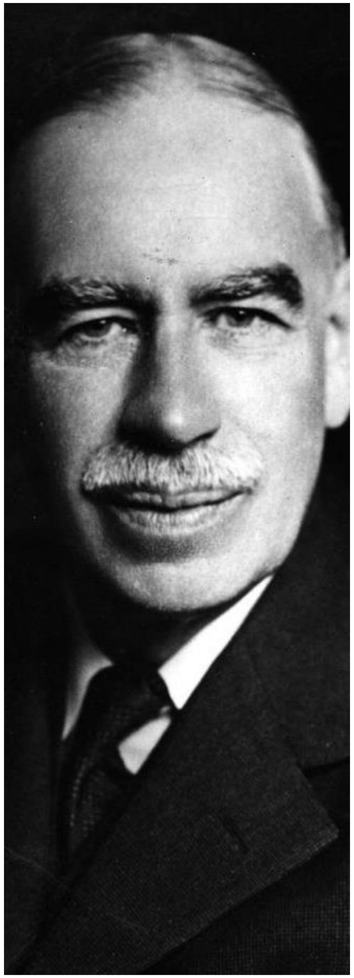
2.5.2 Sex

- Automated “sex workers”?
- Human attachment to machinery – or true friendship (Danaher 2019 vs Nyholm & Frank) and feeling?
- Corruption of humans? Continuation of slavery and prostitution?

2.6 The effects of automation on employment

- Productivity through automation → Job loss?
  - Farming 60% of workers in 1800, <5% now
- Usually: Job loss → productivity growth → job gain
  - (i) the nature of interactions between differently skilled workers and new technologies affecting labour demand and (ii) the equilibrium effects of technological progress through consequent changes in labour supply and product markets. (Goos 2018: 362)
- … is it different this time?
  - Automating blue vs. white collar workers
  - 'dumbbell' market?





Distributive Justice (winners & loosers)

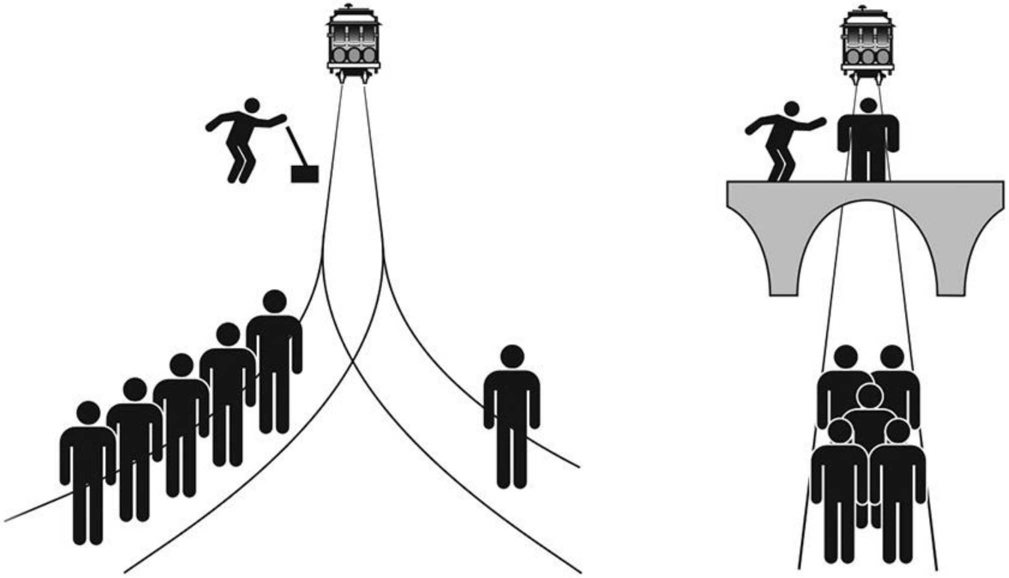
- J. M. Keynes 1930: *Age of Leisure* in 100 years at 1% growth
  - What went wrong??
- Distributive justice (fairness)
  - Decide distribution behind a “veil of ignorance” (J. Rawls 1971) - as if one does not know what position in a society one would actually be taking (labourer or industrialist, etc.)
- Distributive justice in the AI & IT industry?
  - Largely unregulated
  - Winner-takes-all markets
  - Intangible assets (”capitalism without capital”)
- PS: How about the environment??

2.7 Autonomous Systems

- 2.7.1 Autonomy Generally
  - A system is autonomous relative to *x*, to a degree *y*, in pursuing its goals (e.g. to human control to a certain degree)
  - more autonomy → less control & more interaction
  - → who is responsible?
- 2.7.2 Example a) Autonomous Vehicles
- 2.7.3 Example b) Autonomous Weapons

2.7.1 Autonomous Vehicles

- Utility gains (1M deaths/y)
- Distribution of risk & responsibility
- Trolley Problems?



2.7.2 Autonomous Weapons

- Lethal autonomous weapon systems (AWS or LAWS) – tanks, ships, drones, submarines, …
  - Take responsibility away from humans
    - Support extrajudicial killings/war crimes
  - Threaten human dignity?
  - Make wars or killings more likely
- Dystopia or Utopia?

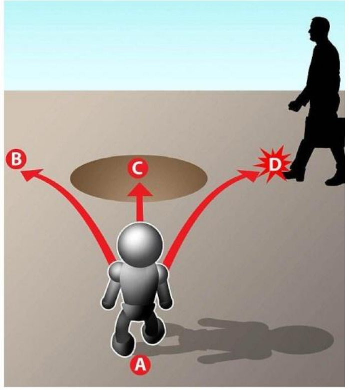
(https://philpapers.org/archive/MLLAKR.pdf)

2.8 Machine Ethics

- What should the machine do? vs. What should the human do? – machines as the subjects of ethics = machine ethics

- "… machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable." (Anderson and Anderson 2007: 15)

- *I. Asimov's Laws (1942)*
  1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



Winfield, Alan; Michael, Katina; Pitt, Jeremy and Evers, Vanessa (eds.) (2019), *Machine ethics: The design and governance of ethical AI and autonomous systems* (Proceedings of the IEEE, 107/3)

Classical Machine Ethics

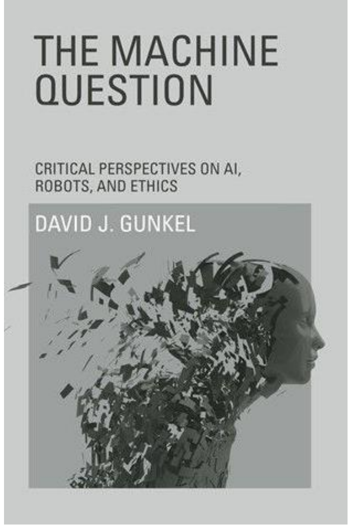
Jim Moor (2006) distinguishes four types of machine ethics:

- ethical impact agents (example: robot jockeys)
- implicit ethical agents (example: safe autopilot)
- explicit ethical agents (example: using formal methods to estimate utility)
- full ethical agents ("can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent.")

→ Is there a machine ethics?

1. Ethics for design & use of machines (e.g. "gun ethics")
2. Ethics for autonomous moral agents (AMA)

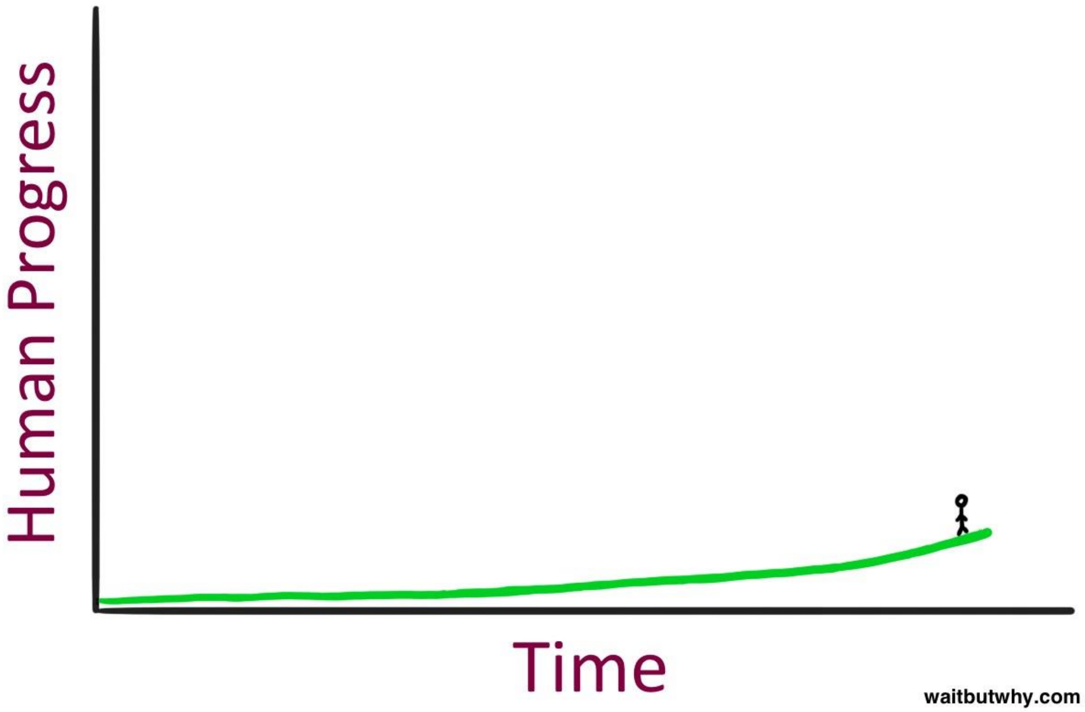
2.9 Artificial Moral Agents



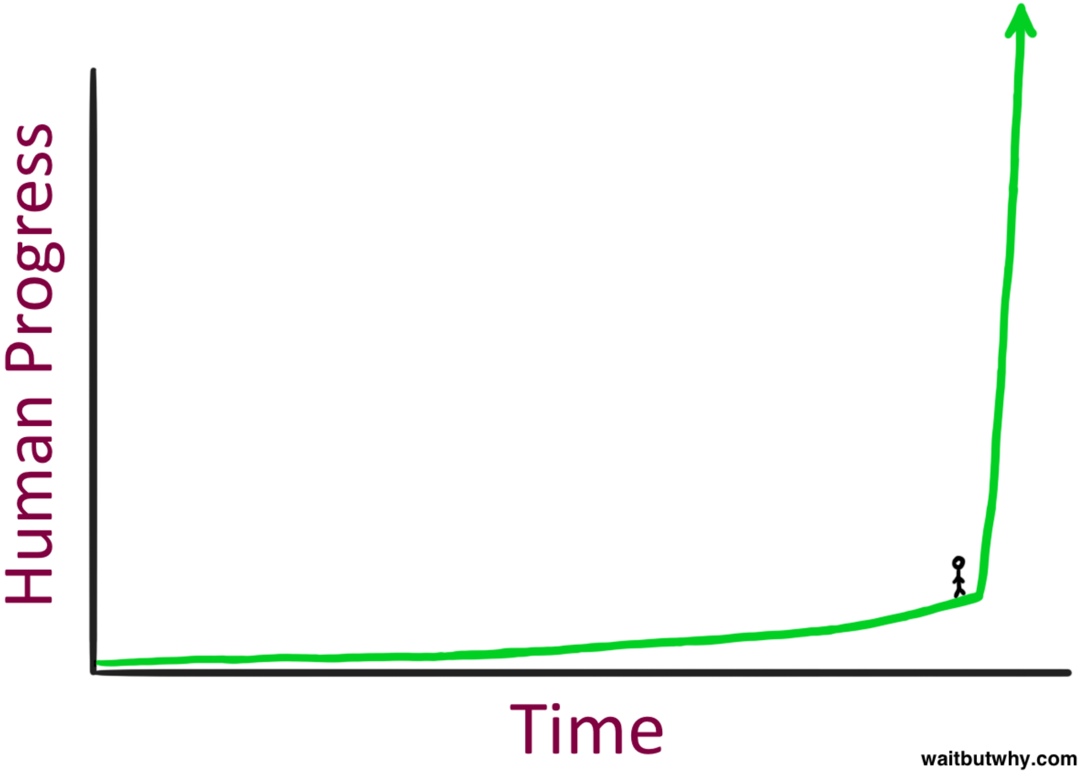
- Responsibility for Robots?
- Rights for Robots?

- What is the basis of attributing moral status & responsibility to systems, is it a set of necessary and sufficient properties?

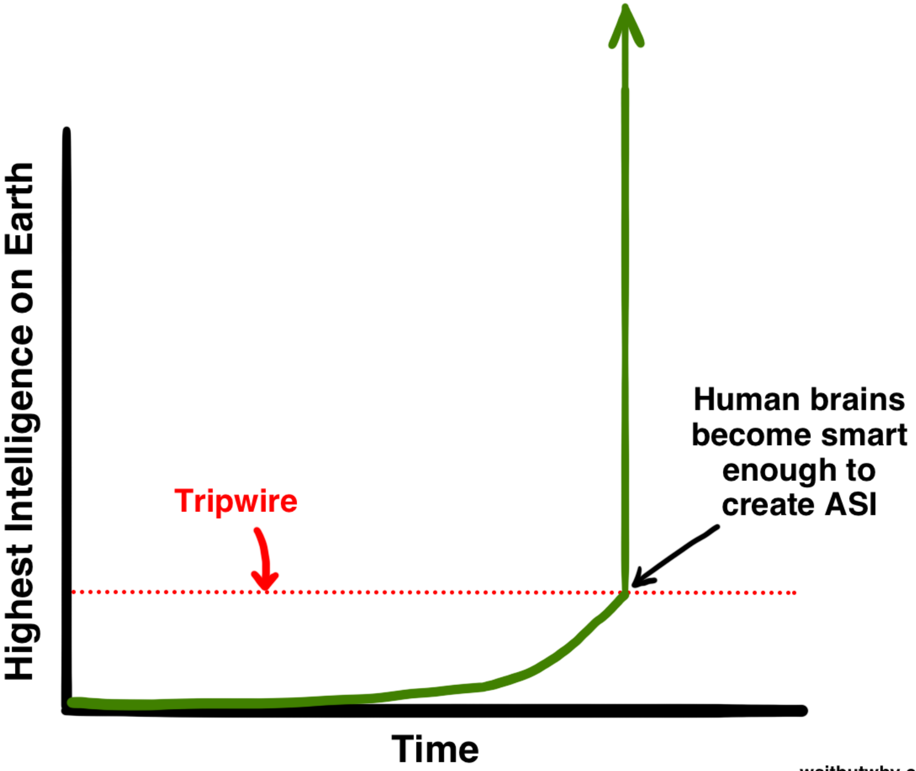
2.10 Singularity … first approximation



waitbutwhy.com

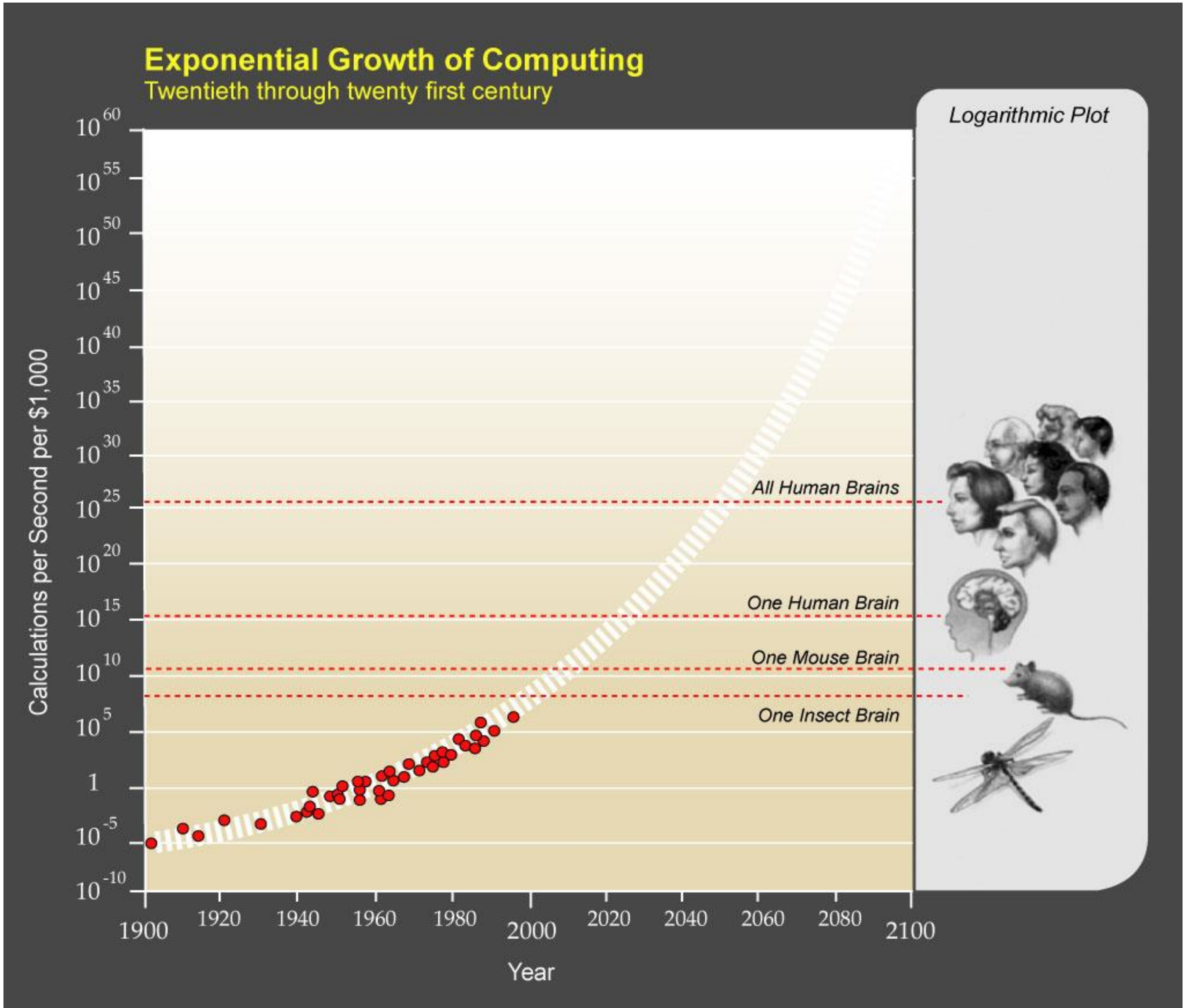

Human Progress
Time
waitbutwhy.com

“Intelligence Explosion”



waitbutwhy.com

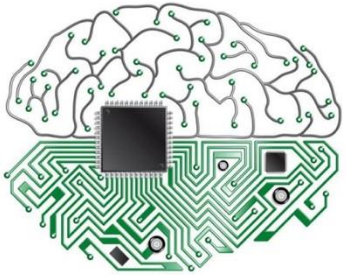
Superintelligence & “Singularity”



Kurzweil 2005, 70

2.10 Singularity

- “Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control“ (Good 1965)
  - -> Existential risk for humanity

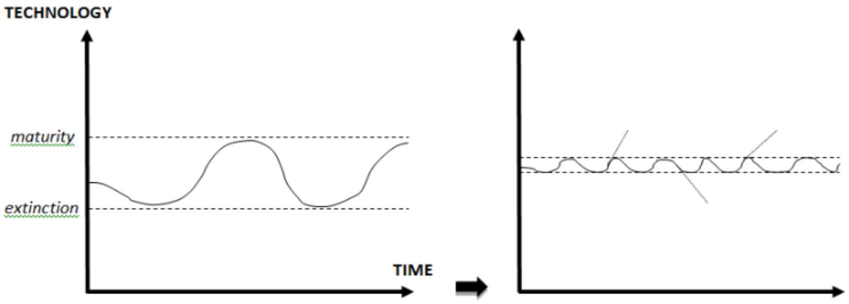


- Issues
  - Will the singularity occur?
  - Is intelligence one-dimensional?
  - Can we hard-wire morality into the system, or control it?
  - Can we know about superintelligence?

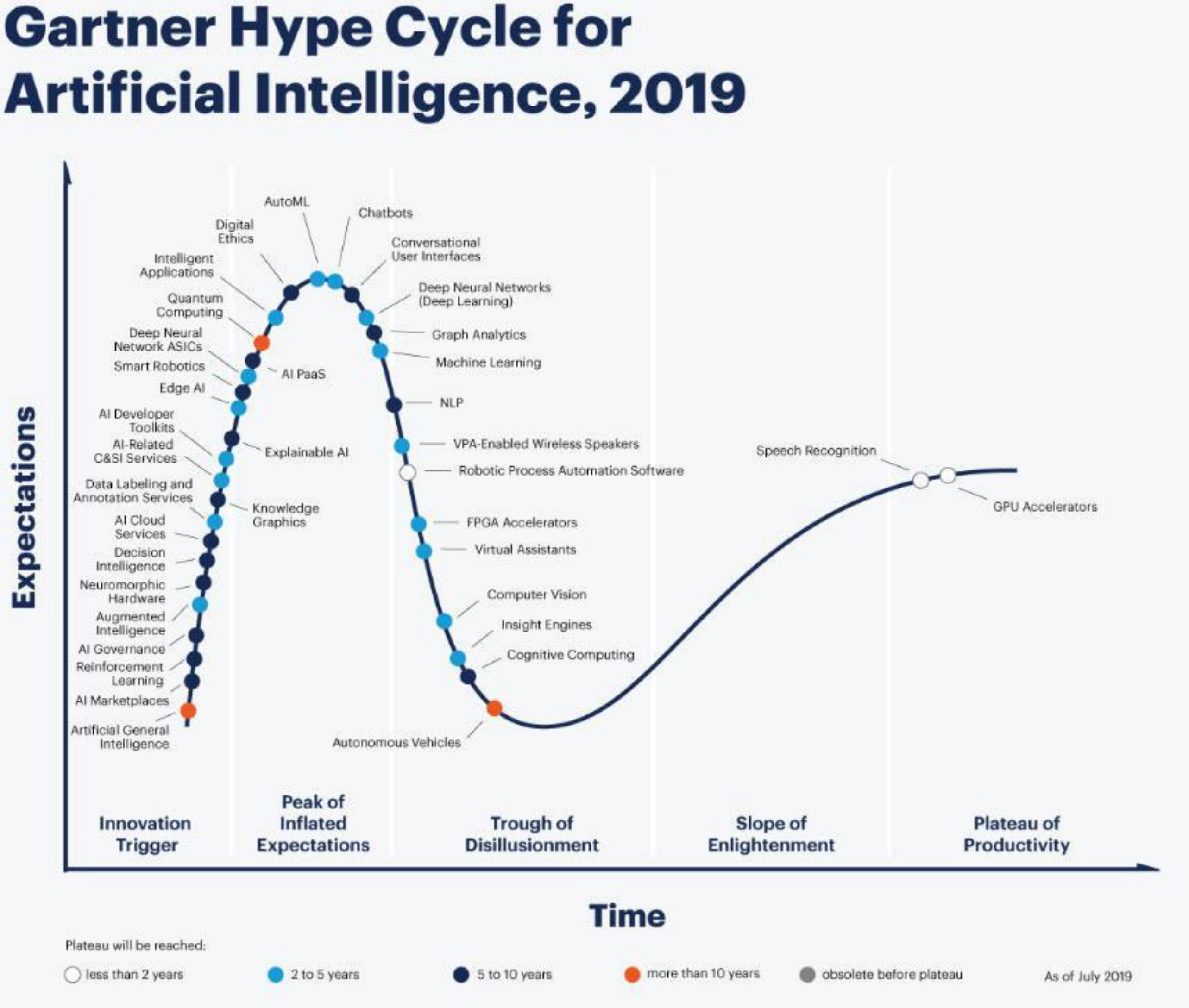
2) Singularity → XRisk

- Superintelligence = "intellects that greatly outperform the best current human minds across many very general cognitive domains" Bostrom 2014, 52

- Superintelligence → Singularity → Existential risk for humanity



4) Where are *we* on the 'hype cycle'? –*AI & AI ethics…*

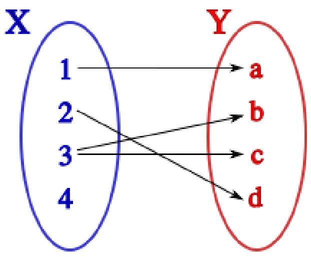


Gartner Hype Cycle for Artificial Intelligence, 2019

“Five innovation profiles debuted on the hype cycle 2020:

- **Composite AI** refers to the combination of different AI techniques to achieve the best results.
- **Generative AI** is the next frontier compared to the AI methods that directly extract numeric or categorical insights from data.
- **Small Data** as a concept indicates both the issue and approach to help those clients who ask us, “How should we get data for AI if we are not Google?”
- **Responsible AI:** The broader AI adoption is, the more enterprises learn about their responsibility for the AI solutions and technologies they implement.
- **Things as Customers:** Customer experience is at the top of corporate AI agendas.”

4) AI Policy
a) A Mapping Problem



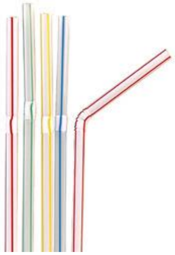
Policy Aims

- Secure Privacy
- Prevent Manipulation
- Prevent Opacity and Bias
- Safe Interaction with Machines
- Safe Autonomous Systems
- Save the World

Forms of Policy

- Law
- Regulation
- Taxation
- Public organisation action
- Stakeholder action
- Principles
- Good-will statements

b) “Policy” examples …



1. Outlawing plastic drinking straws
2. Obligatory recycling of plastic straws
3. Extra tax on one-way plastic straws
4. Obligatory additional price on plastic straws (or bags)
5. Provide natural straws for free
6. Train employees on environmental issues
7. Public information on environmental issues
8. Bottom-up ‘stakeholder’ push for environmental awareness
9. Nudging

Exhibit A:
OECD Principles on AI (May 2019)

1. Inclusive growth, sustainable development and well-being
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability

17 Sustainable Development Goals (United Nations 2015): (1) No Poverty, (2) Zero Hunger, (3) Good Health and Well-being, (4) Quality Education, (5) Gender Equality, (6) Clean Water and Sanitation, (7) Affordable and Clean Energy, (8) Decent Work and Economic Growth, (9) Industry, Innovation and Infrastructure, (10) Reducing Inequality, (11) Sustainable Cities and Communities, (12) Responsible Consumption and Production, (13) Climate Action, (14) Life Below Water, (15) Life On Land, (16) Peace, Justice, and Strong Institutions, (17) Partnerships for the Goals.

Exhibit B: ‘High Level Expert Group’ on AI, EU (April 2019 & 2020)

- Trustworthy AI:
  1. lawful
  2. ethical
  3. technically robust

- Requirements for trustworthy AI:
  1. human oversight
  2. technical robustness
  3. privacy and data governance
  4. transparency
  5. fairness
  6. well-being
  7. accountability

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

AI

ETHICS GUIDELINES
FOR TRUSTWORTHY AI

Exhibit C:
EC regulatory proposal (April 2021)

“Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts” (April 2021)

- “The general objective of the intervention is to ensure the proper functioning of the single market by creating the conditions for the development and use of trustworthy artificial intelligence in the Union “

- The regulation follows a risk-based approach, differentiating between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk.

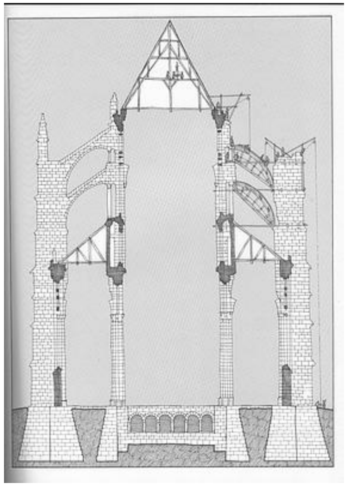
Ongoing initiatives …

- 2020-: Global Partnership on AI (GPAI), G7 + EU …
  - ≠ “Partnership on AI” (Amazon, Facebook, Google, DeepMind, Microsoft, IBM, Apple, Baidu …)
- 2020-: OECD Network of Experts on AI (ONE AI)
- …
  - [http://www.pt-ai.org/TG-ELS/policy]

4) “Ethics of AI & Robotics”
What’s hot & what’s cold?

- **Criteria**
  - Theoretical interest (‘ethical problem’)
  - Practical interest (utility, injustice)

2. **Main Debates**



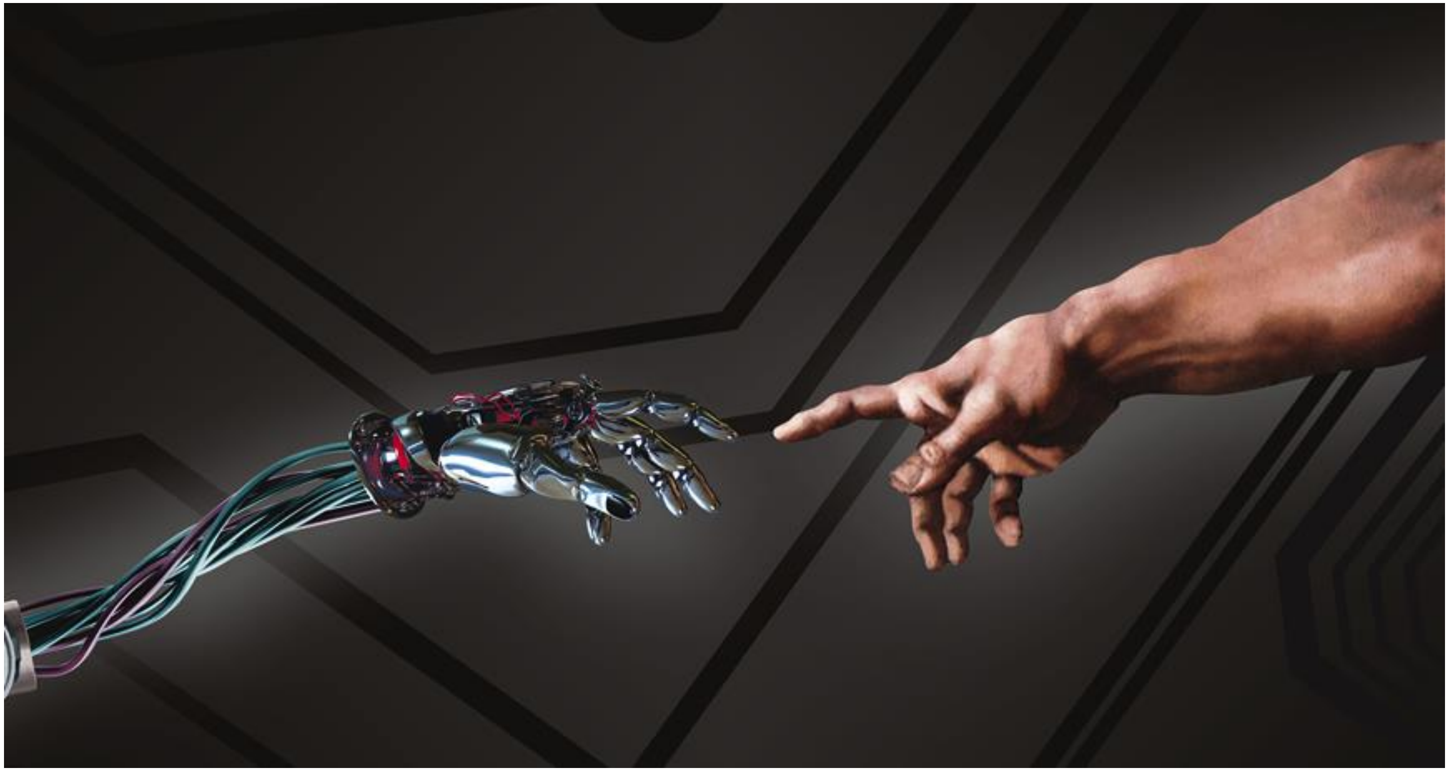
2.1-2 Data: Privacy & Manipulation

2.3-4 Epistemology: Opacity & Bias

2.5-7 Robot Ethics: Automation, Interaction, Autonomy

2.8-9 Concepts (Agency, Responsibility, Autonomy …)

2.10 Singularity



Thank You!

PS:
Three dimensions of privacy

- Decisional privacy
- Informational privacy
- Local privacy

→ All three dimensions serve to protect individual freedoms.

Source: Beate Rössler (2018), “Three Dimensions of Privacy”, in: B. van der Sloot and A. Groot (eds.), *The Handbook of Privacy Studies. An Interdisciplinary Introduction*, Amsterdam: AUP, 137-142. 0r

Roessler, Beate (2008) New Ways of Thinking about Privacy. In: Dryzek, J., Honig, B., Phillips, A., (eds.) The Oxford Handbook of Political Theory. Oxford: Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199548439.001.0001/oxfordhb-9780199548439-e-38

Decisional privacy [autonomy]

- Core: idea of physical privacy (privacy of the **body**)
  - → Physical privacy is necessary in order to live a self-determined and authentic life.
  - → core: who you want to be and how you want to live.
- The right to abortion was grounded by an appeal to a right to privacy. Judgement of the US Supreme Court in the case *Roe vs. Wade* (1973)
- freedom in social spaces and in relation to other individuals
- respect for decisional privacy → expectations of non-interference and indifference

Informational privacy [data privacy]

- Who knows what about me? Who has which information?
- Becomes more and more important in our digital society, but also more and more difficult to protect
- Deeply connected with autonomy
  - → control over our own self-representation and over what others know about us, and to use this control to regulate our relationships and thus the different social roles that we play.
  - Serves the protection *of* special relationships and *within* relationships
  - Challenge: Smart City Eindhoven: constant data flow lets it optimise services constantly → Who owns all the data produced by the city of the future? Who controls it? Whose laws apply?

Local privacy

- Privacy of my space, the home
- Gives me/us the possibility to be alone and "to be ourselves"
- Problematic from a feminist perspective? (Room for oppression of women & others)

The value of privacy – intrinsic or instrumental?

Intrinsic

- Privacy is valuable in itself, independently of any valuable ends that it can be a means to.
- Source: a form of respect that we owe to all rational beings (Kant)

Instrumental

- Privacy has value as a means to achieving an end that is valuable.
- Privacy is valuable, for example, as a means to strengthen a person’s autonomy or as a means to protect intimate relationships.