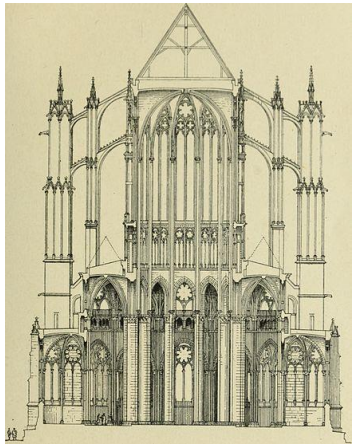
Ethics & Policy of Artificial Intelligence and Robotics

Vincent C. Müller

TU Eindhoven / U Leeds / Alan Turing Institute
www.sophia.de
12.06.2019

A) Structure: AI Ethics

(forthcoming: Stanford Encyclopedia of Philosophy)



1. Introduction

1.1. Background of the Field

1.2. AI & Robotics

1.3. A Note on Policy

2. Ethics for the Use of AI & Robotics Systems

2.1. Privacy, Surveillance & Manipulation

2.2. Our Epistemic Condition: Opacity and Bias

2.3. Interaction with Machines

2.4. The Effects of Automation on Employment

2.5. Autonomous Systems

3. Ethics for AI & Robotics Systems

3.1. Machine Ethics

3.2. Artificial Moral Agents

4. Singularity

4.1. Singularity and Superintelligence

4.2. Existential Risk from Superintelligence

4.3. Controlling Superintelligence?

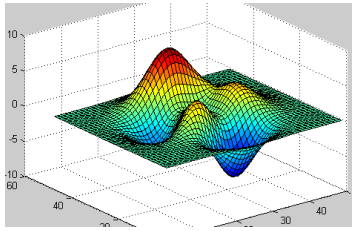
1.2 AI & Robotics

- There is no "ethics of AI and robotics"
  - *-> map problems and discussions/positions*
- AI: any kind of machine that shows intelligent behaviour, i.e. complex behaviour that is conducive to reaching goals (based on computation)
  - Classical AI, cognitive science, machine learning
- Robots: physical machines that have actuators that interact with the environment, such as a gripper or a turning wheel

2.1 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)
- Manipulation of action beyond economic aims
- Manipulation of text, images, video, …

2.2 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

2.3.2. Care

- Robots in health care - de-humanising care?
- E.g. lifting patients, transporting material, eating with robot arm, robots for comfort
- Can there be ‘care robots’? The dystopia is non-care.

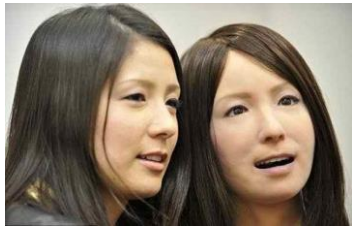
2.5.3 Autonomous Weapons

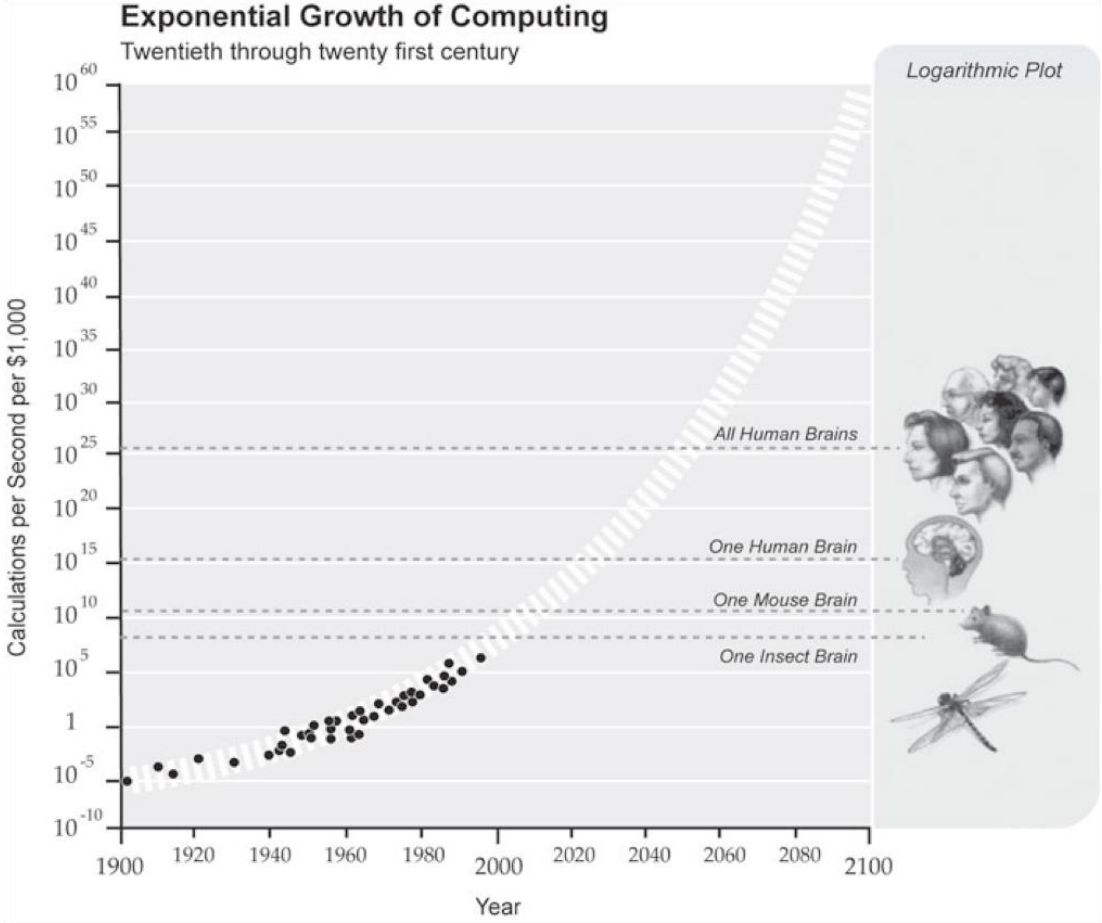
- Lethal autonomous weapon systems (AWS or LAWS) – tanks, ships, drones, submarines, …
  - Support extrajudicial killings/war crimes
  - Threaten human dignity
  - Take responsibility away from humans
  - Make wars or killings more likely
- Dystopia or Utopia?

3.2. Responsibility for Robots

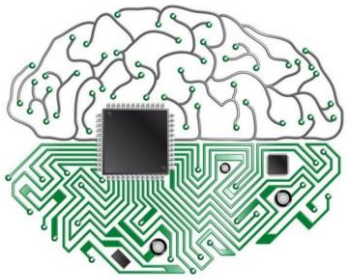
- Can the robots themselves be responsible, liable or accountable for their actions?
- Robots as moral agents, as moral patients?
- Sentience & artificial consciousness
- Should the distribution of risk should take precedence over discussions of responsibility?

4.3 Machine Ethics for Superintelligence - The Control Problem



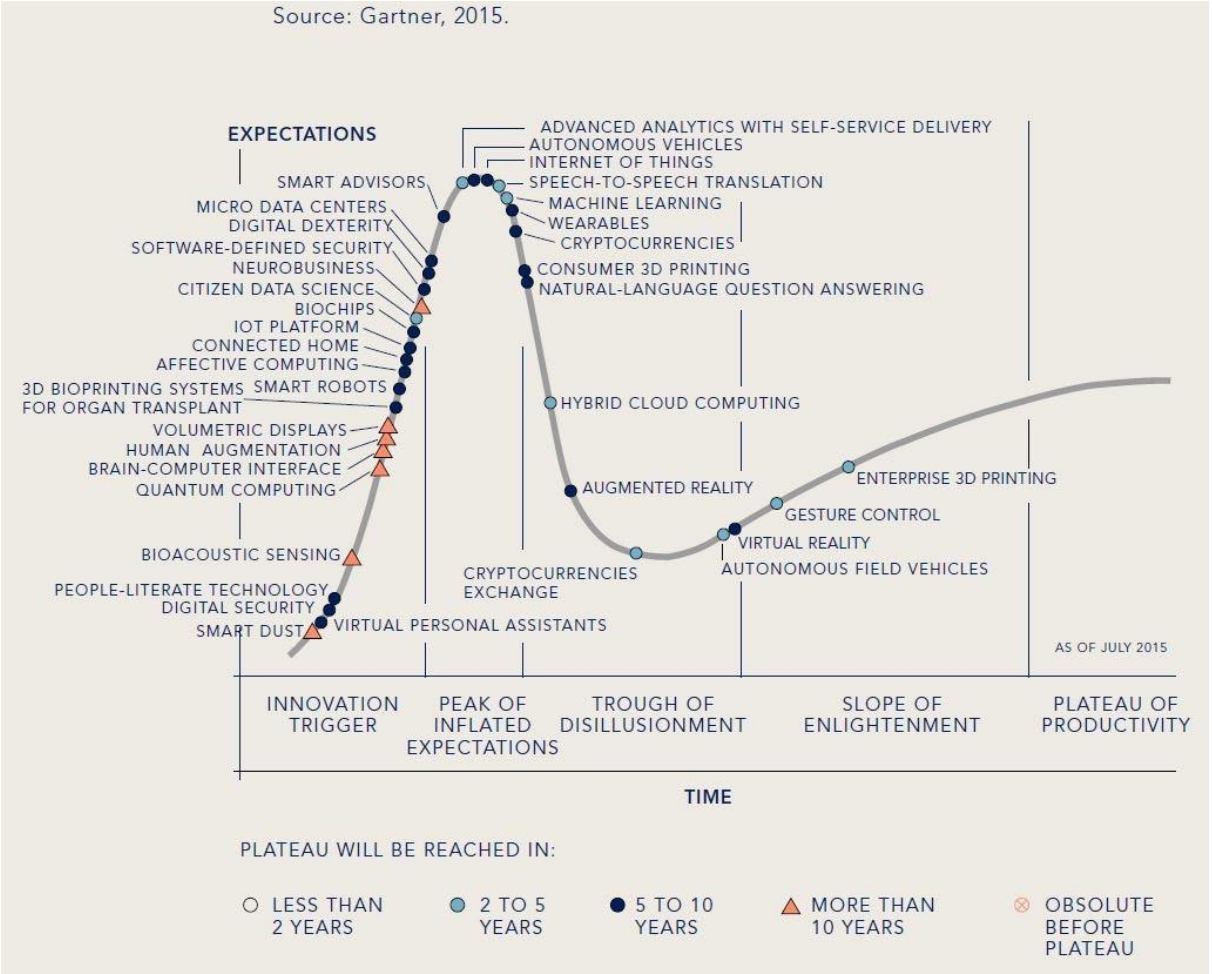


Kurzweil 1999



- "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?

Where are we on the 'hype cycle'?

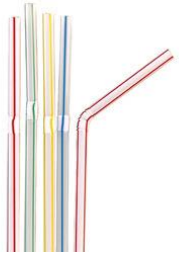


B) Structure: Policy

- Shapes and Forms
  - Law, Regulation, Self-regulation, Incentives, finances, …
- Where do we need policy? Where do we need new policy?
- Do we know what policy should achieve?

[a list on http://www.pt-ai.org/TG-ELS/policy]

"Policy" examples (environmental – in biotech?)



1. Outlawing plastic straws
2. Obligatory recycling of plastic bottles
3. Tax on one-way takeaway food packaging
4. Obligatory price on plastic shopping bags
5. Higher tax on gasoline + no tax on electric vehicles
6. Training employees on environmental issues
7. Employees get a bicycle, but no car parking space
8. Public information on environmental issues
9. Bottom-up 'stakeholder' push for environmental awareness
10. Private purchasing decisions

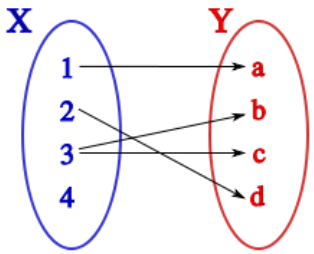
Exhibit A: OECD Principles on AI (May 2019)

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

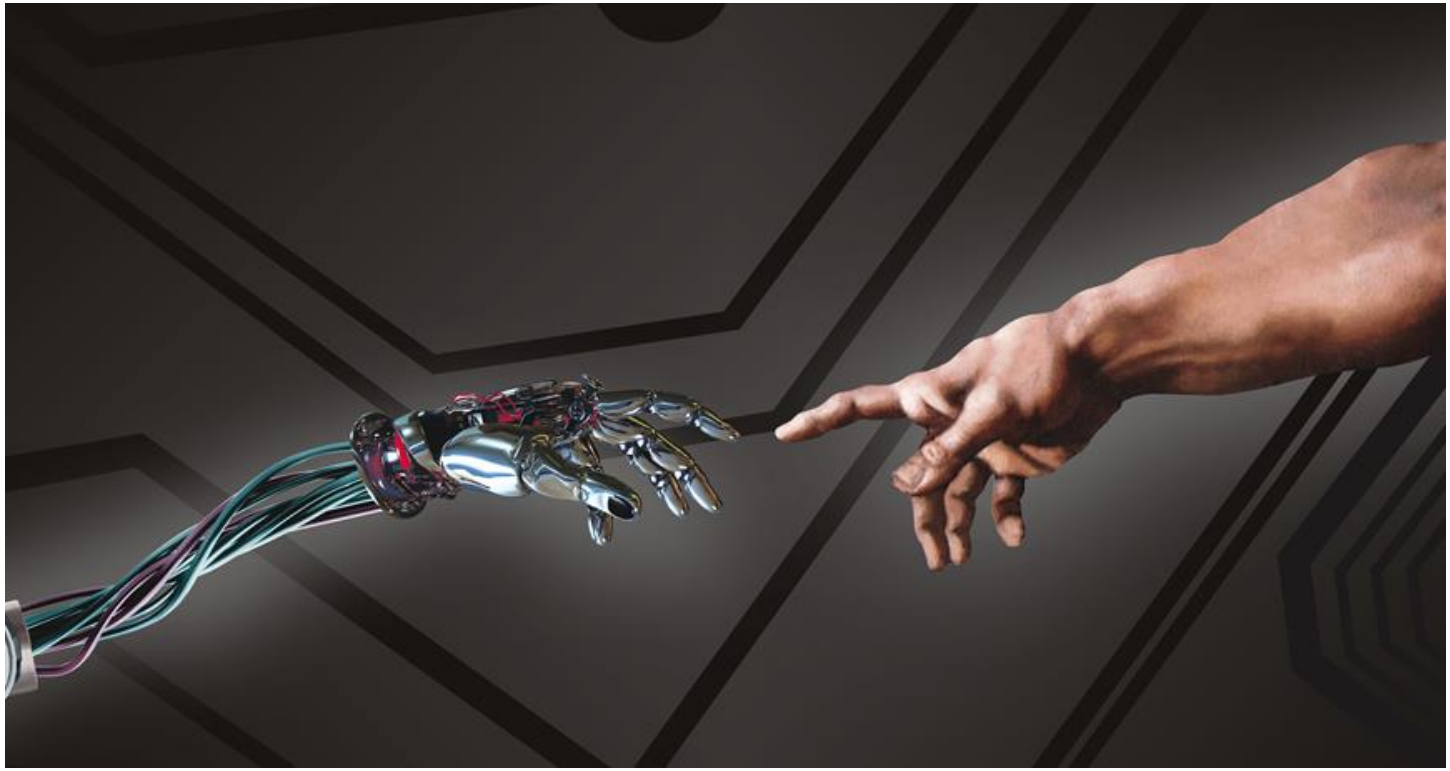
Exhibit B: AI 'High Level Expert Group', EU (April 2019)

- 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

AI Ethics & Policy
A Mapping Problem



- 2.1. Privacy, Surveillance & Manipulation
- 2.2. Our Epistemic Condition: Opacity and Bias
- 2.3. Interaction with Machines
- 2.5. Autonomous Systems
- 4.3. Controlling Superintelligence?
- Law
- Regulation
- Taxation
- Public organisation action
- Stakeholder action
- Principles
- Good-will statements



Thank You!