Meaningfull Human Control
Designing for Ethical AI

Catholijn M. Jonker
c.m.jonker@tudelft.nl

Universiteit Leiden

TUDelft

Netherlands Organisation for Scientific Research

The automation perspective on AI

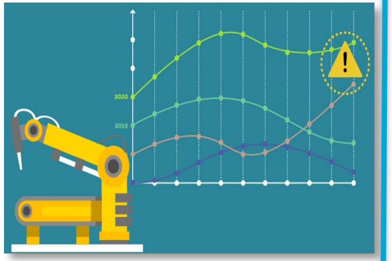
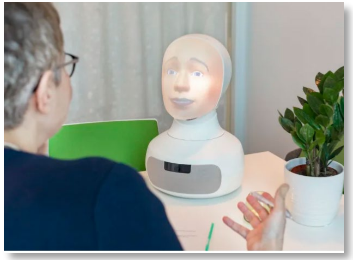
“My guess for when we will have full autonomy [in cars] is approximately three years” (Elon Musk, 2015)

“[a] highly-trained and specialised radiologist may now be in greater danger of being replaced by a machine than his own executive assistant” (Andrew Ng, The Economist, 2016)

“People should stop training radiologists now. It’s just completely obvious that within 5 years, deep learning is going to do better than radiologists”
(Geoffrey Hinton, The New Yorker, 2017)

Hi

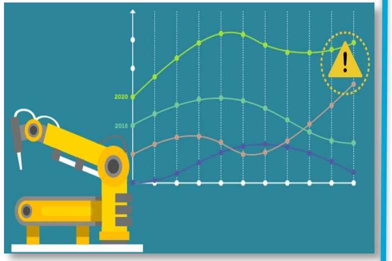
AITECH

	Predictive maintenance	Life expectancy prediction	Job suitability prediction	Sexual orientation prediction
				
Criterion/measure	Repair costs	Risk score	Avoid hiring unsuitable	Reproduce human ability
features	* vibration spectrum * resource usage * current signature	* blood pressure * creatine * white blood cell count	* facial emotion * voice timbre * vocabulary	* facial features * morphology * grooming
optimiser optimised	Maintenance planning Reduced value of expertise	Informed treatment decisions Root cause unclarity	High volume selection Reduced self presentation	Security, marketing Stigmatization

TUDelft

We are not taking a moral stance. The issue is that you inspect your own arguments.

Predictive maintenance

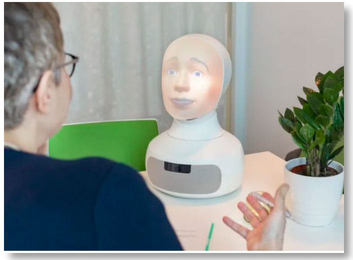


Repair costs

Life expectancy prediction

Risk score

Job suitability prediction



Avoid hiring unsuitable

Sexual orientation prediction

Reproduce human ability

- Moral alignment is missing.
- Is this legally acceptable?
- Predictive analytics decides your fate in a non-forgiving way.
- Where are the checks and balances?

- Borderless innovation without care seems irresponsible.
- Not in the anthropomorphic way, human-like vengeance, more in simple mistakes that we should or should not have made.
- **What causes these ‘simple mistakes’?**

Design & engineer it better

- Shift from autonomous AI to Hybrid Intelligence
- Improve AI’s situation awareness
- Raise ethical awareness in humans & AI
- Place AI under meaningful human control

The future of human kind with Artificial Intelligence

- Empower autonomy
- Expand experience
- New activities
- Strenghten democracies

- Reduce autonomy
- Replace experience
- Redundant
- Endanger democracies

The future of human kind with Artificial Intelligence

- Empower autonomy
- Expand experience
- New activities
- Strenghten democracies

- ~~Reduce autonomy~~
- ~~Replace experience~~
- ~~Redundant~~
- ~~Endanger democracies~~

Align AI with human & social values
& create Hybrid Intelligence

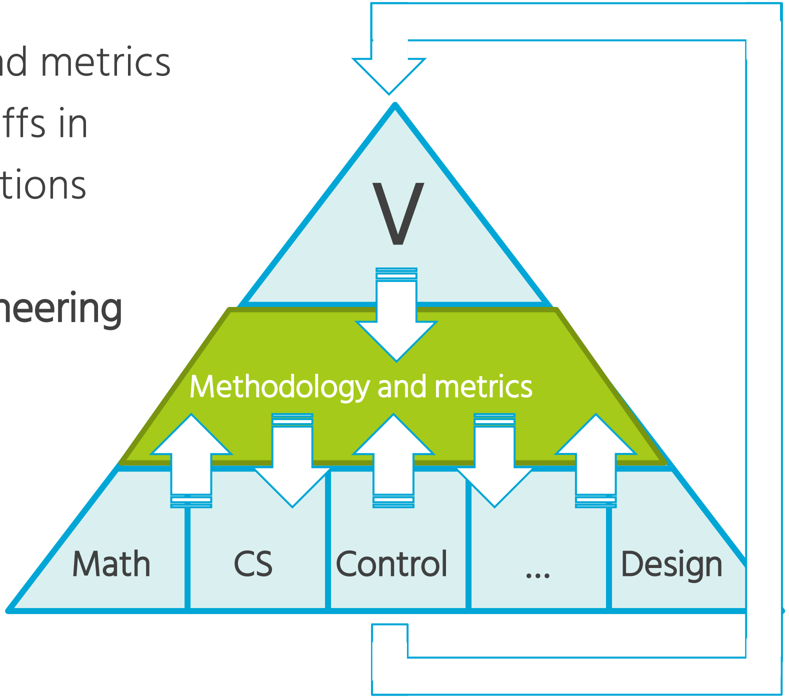
An engineering perspective

Methodology and metrics to guide trade-offs in component solutions

Design and engineering research

Iterative

Multidisciplinary



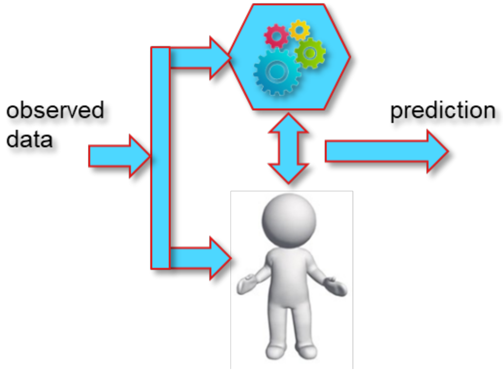
Values meets engineering

Meaningful human control in AI:

- Tracking
- Tracing

Meaningful Human Control

Humans must be in a **position** to | be **capable** of | **expected** to ‘take over control of’ the system



Synergistic & transparent collaboration

Create awareness & ownership

Develop definitions & quantifiable criteria

Engineer concrete Hybrid Intelligence

Meaningful human control

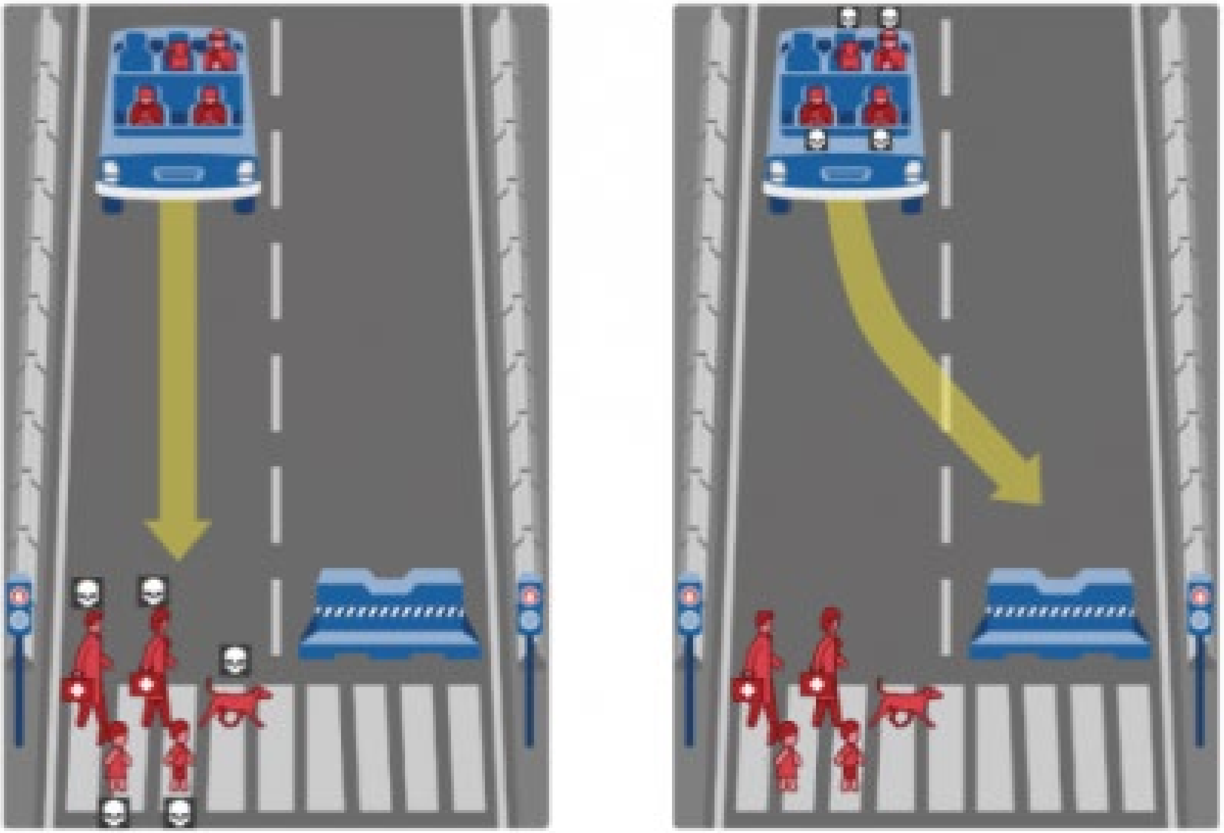
- it is important to educate
- we need to understand it
- we build and break it

Leadership, control & responsibilities in teams

- Shared leadership is
  - a social process that requires team leadership from team members
- Flowing leadership & responsibilities
  - From one team member to another
  - Depending on the situation
  - Depending on the capabilities of the team members
- **Shared control**

moralmachine.mit.edu

Teach it ethics



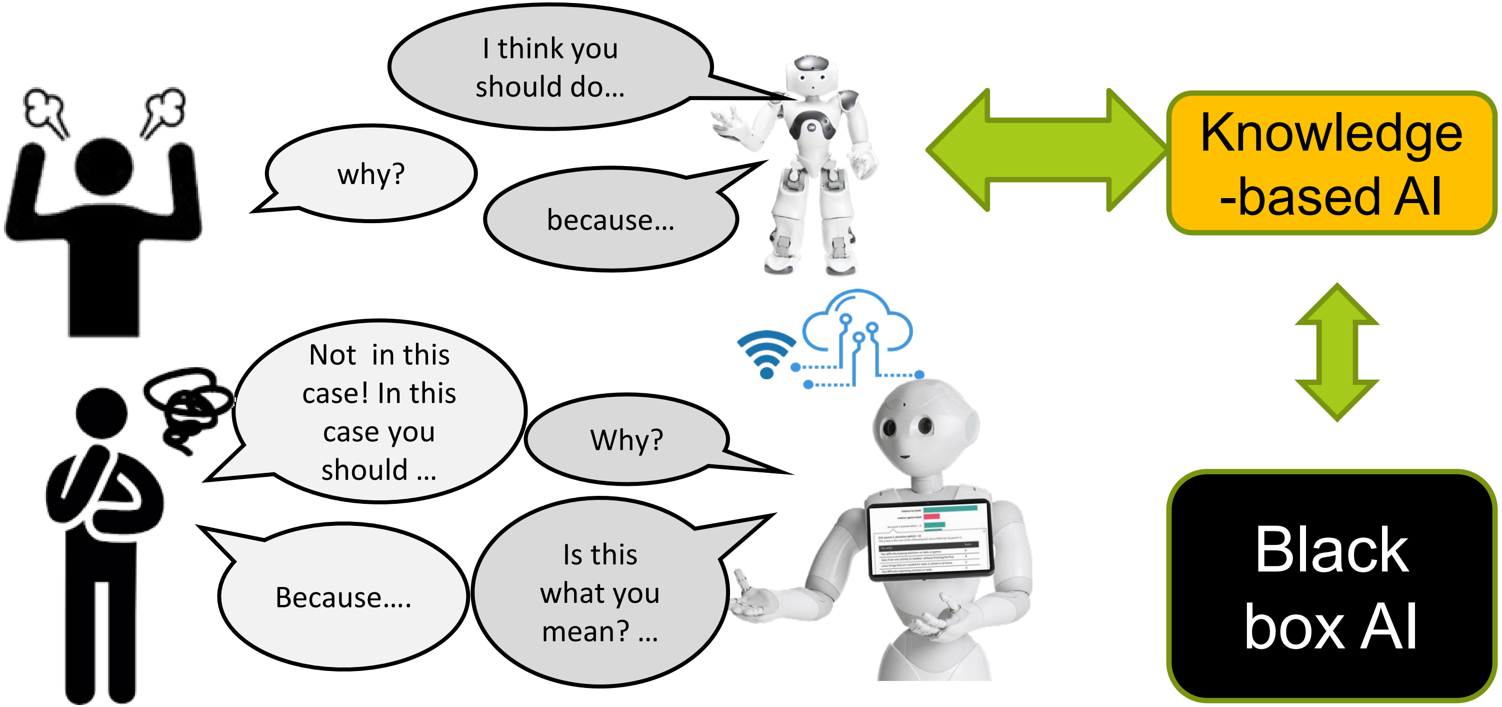
Ethical Autonomous Vehicle

- Utilitarian car (Teleology)
  - The best for most; results matter
  - maximize lives

- Kantian car (Deontology)
  - Take no harmful action; people matter
  - do not take explicit action if that action causes harm
- Aristotelian car (Virtues Ethics)
  - Pure motives; motives matter
  - Harm the least; spare the least advantaged (pedestrians?)

Can you personalise yours? Is that ethical?

Hybrid Intelligence over AI

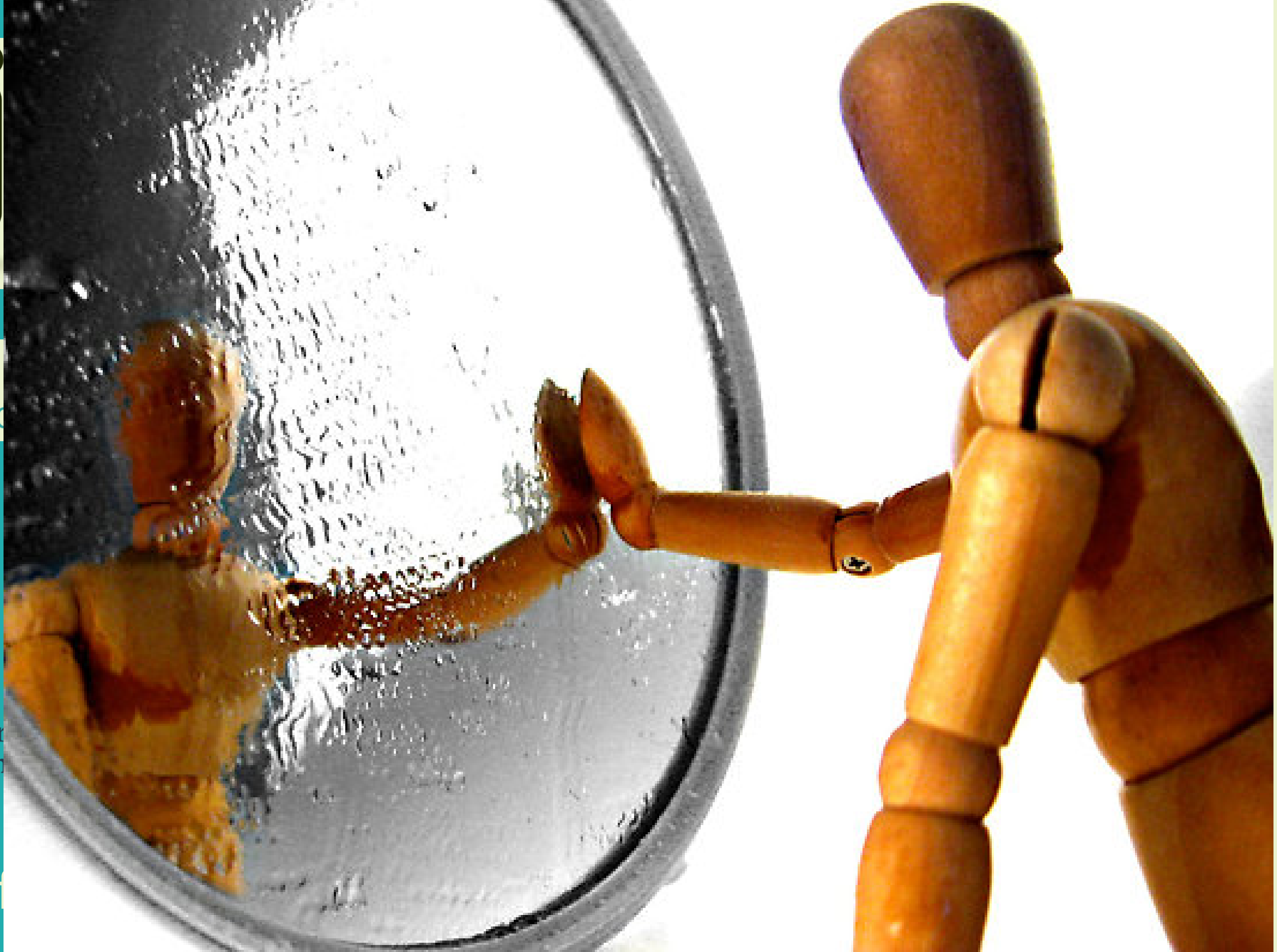


Knowledge-based AI

Black box AI

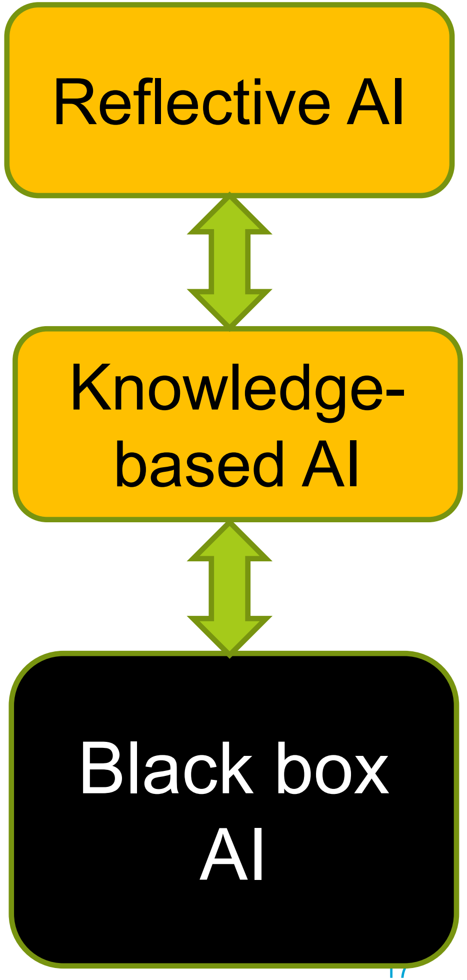
Mission

- Shift from autonomous AI to Hybrid Intelligence
  - Replace → Augment
  - Autonomy → Co-activity
  - Isolated AI → Escalate to HI
  - Hybrid Intelligence → System of Hybrid Systems
- Improve human & AI’s situation awareness
- Raise ethical awareness in humans & AI
- Place AI under meaningful human control



Self-reflective Hybrid Systems

- Where are we from a moral point of view?
- What biases are we forming?
- What is the quality of the data we use?
- Who has the expertise we need?
- Epistemic logic:
  - What do we know?
  - What do we know that we don't know?
  - Unknown Unknowns



Self-reflective Hybrid Intelligent systems: combining the strenghts of

Machine Learning

Knowledge Representation

Human Intelligence

https://www.hybrid-intelligence-centre.nl/
https://www.delftdesignforvalues.nl/
https://www.tudelft.nl/aitech

Catholijn M. Jonker