

# ENGINEERING MEANINGFUL HUMAN CONTROL

## HOE REALISEREN WE DE ROL VAN DE MENS IN AUTOMATISCHE (\*AI) SYSTEMEN

Prof.dr.ir. Inald Lagendijk / TU Delft, the Netherlands

# Rekenkamer: nauwelijks aandacht voor ethiek bij algoritmes overheid

26 januari 2021 17:43

Aangepast: 26 januari 2021 18:07

RTLnieuws

Er gaat van alles mis met het gebruik van algoritmes overheid, concludeert de Algemene Rekenkamer in haar rapport. Op tal van ministeries wordt bijvoorbeeld te weinig besteed aan de ethische aspecten van algoritmes, met name discriminatie.

Op slechts drie van alle twaalf ministeries staat een prioriteitenlijst als er wordt nagedacht over een gesprek met de Rekenkamer voerde met meerdere ministeries. Welke ministeries dat zijn, wil de Rekenkamer onafhankelijke controleur van de overheid, over het gebruik van algoritmes was niet het doel van het onderzoek, legt een woordje uit.

Het doel was vooral om te kijken hoe het gebruik van algoritmes overheid is geregeld. En dat is niet best, blijkt. Op veel van de ministeries vaak niet eens welke algoritmes er gebruikt worden.

## Wat is een algoritme?

Een algoritme is een verzameling regels die een computer op grote schaal taken te verrichten die mensen kunnen doen. Zo kan een algoritme bijvoorbeeld gebruikt worden om te sporen, door allerlei gegevens door te spitten en te wijzen.

## Geen overzicht

Pas na het belletje van de Rekenkamer vanwege het gebruik van veel ministeries en bij CIO Rijk een lampje branden over welke algoritmes er eigenlijk al gebruikt worden.

## Rekenkamer: ministers weten niet welke algoritmes ze gebruiken

Jan Fred van Wijnen 26 Jan 

### Discriminatie op de loer

De Rekenkamer heeft ook gemerkt dat ambtenaren soms niet weten of ze een algoritme of een gewoon computerprogramma gebruiken.

Algoritmes zijn programma's die zelfstandig data verwerken. Daardoor kunnen ze een beslissing nemen bij een relatief simpele subsidie-aanvraag. Maar ze kunnen ook voorspellen welke burgers mogelijk frauderen. Daar sluipt makkelijk discriminatie in van etnische bevolkingsgroepen.

Dit gebeurde bij het programma Syri, dat werd gebruikt door de Belastingdienst en het UWV om uitkeringsfraude te voorspellen. Het algoritme is ontwikkeld met gegevensbestanden waarin relatief veel allochtone burgers voorkomen. Daardoor 'leerde' het algoritme dat allochtonen eerder frauderen dan niet-allochtonen. Syri werd in februari 2020 door de rechter verboden, na een rechtszaak die burgers hadden aangespannen tegen de staat.

### Toets op de 'menselijke maat'

De algoritmes die nu zijn onderzocht, zijn geen programma's die helemaal zelfstandig kunnen 'leren' van databestanden. Daardoor konden de onderzoekers zien hoe de computer een beslissing neemt. Er is nog geen zogeheten 'black box' aangetroffen.

## Uber's self-driving operator charged over fatal crash

16 September 2020

BBC

NEWS

The back-up driver of an Uber self-driving car that killed a pedestrian has been charged with negligent homicide.

Elaine Herzberg, aged 49, was hit by the car as she wheeled a bicycle across the road in Tempe, Arizona, in 2018.

Investigators said the car's safety driver, Rafael Vasquez, had been streaming an episode of the television show The Voice at the time.

Ms Vasquez pleaded not guilty, and was released to await trial.

Uber will not face criminal charges, after **a decision last year that there was "no basis for criminal liability"** for the corporation.

The accident was the first death on record involving a self-driving car, and resulted in Uber ending its testing of the technology in Arizona.

## 'Visually distracted'

Lengthy investigations by police and the US National Transportation Safety Board (NTSB) found that human error was mostly to blame for the crash.

Ms Vasquez was in the driver's seat, and had the ability to take over control of the vehicle in an emergency.

Dash-cam footage released by police showed Ms Vasquez looking down, away from the road, for several seconds immediately before the crash, while the car was travelling at 39mph (63km/h).

# Digitale Maatschappij

- **Digitalisering:** het gebruik van digitale data en intelligente algoritmen in essentiële maatschappelijke en economische processen, interacties, adviezen en besluitvorming.
- Voorbeelden te over, wekelijks in nieuws.
- In vrijwel elke sector van maatschappij, economie en wetenschap.
- **Voordelen** staan soms op gespannen voet met **risicos en verantwoordelijkheid**.

# Voordelen. Risicos. Verantwoordelijkheid

- **Radiologie.** *Is dit een tumor?*
- **HR afdeling.** *Wat voor een soort CV is dit?*
- **Immigratie.** *Is dit een migrant of vluchteling?*
- **Beurshandelaar.** *Wat is het risicoprofiel van een investering?*
- *Opereren of niet? Welke behandeling?*
- *Uitnodigen voor een interview? Baan aanbieden?*
- *Status? Verblijfsvergunning of niet?*
- *Hoeveel, wanneer te (des)investeren?*

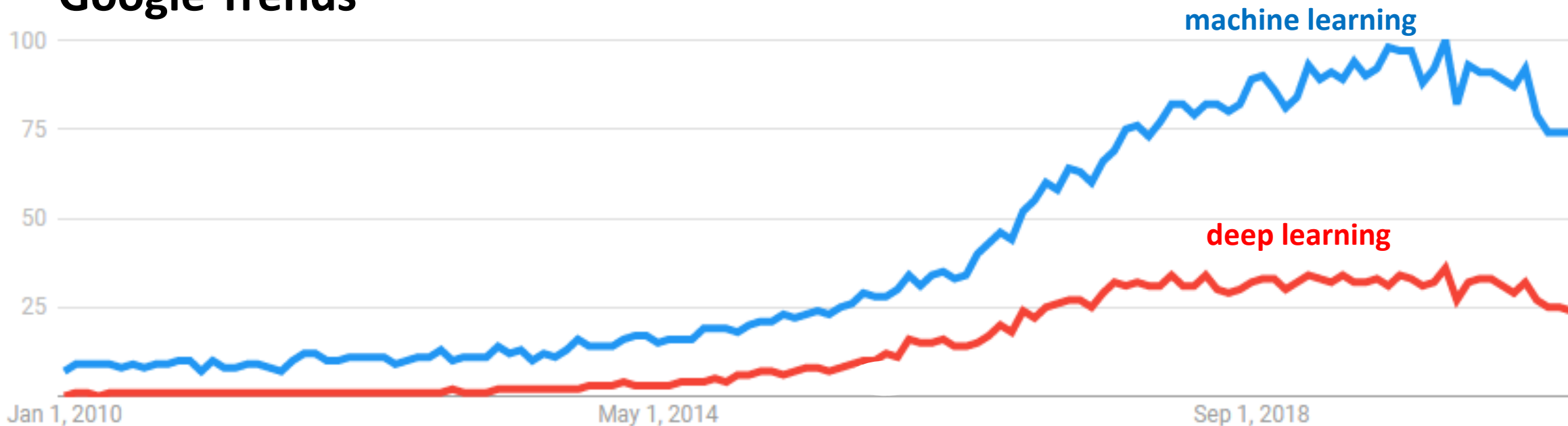
Hoe slim is het (AI) algoritme?

Kan het (AI) algoritme deze verantwoordelijkheden dragen?

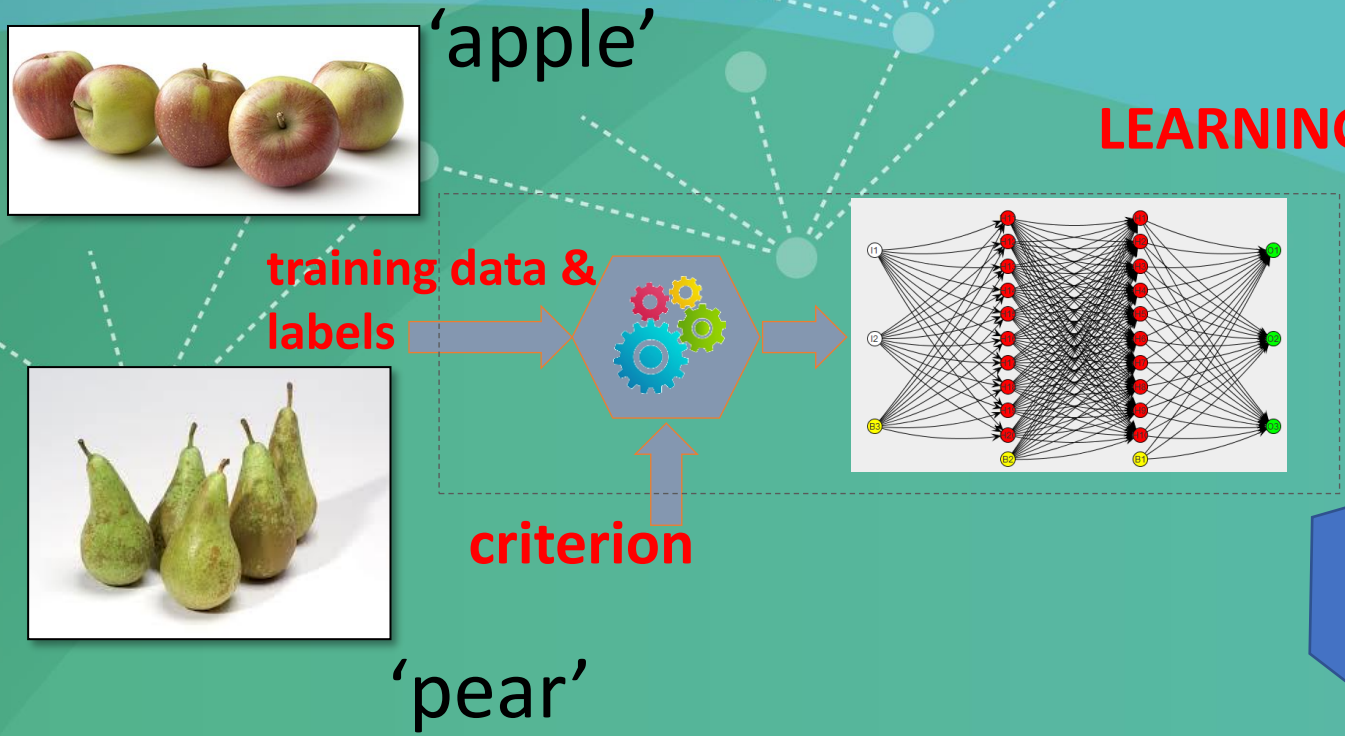


# Machine leren: de AI succes motor

## Google Trends



# Machine leren: de AI succes motor



**Model:**

- 1. De relevanten wereld kennis**
- 2. Relevante criteria voor prestatie**

# Machine Learning



'apple'

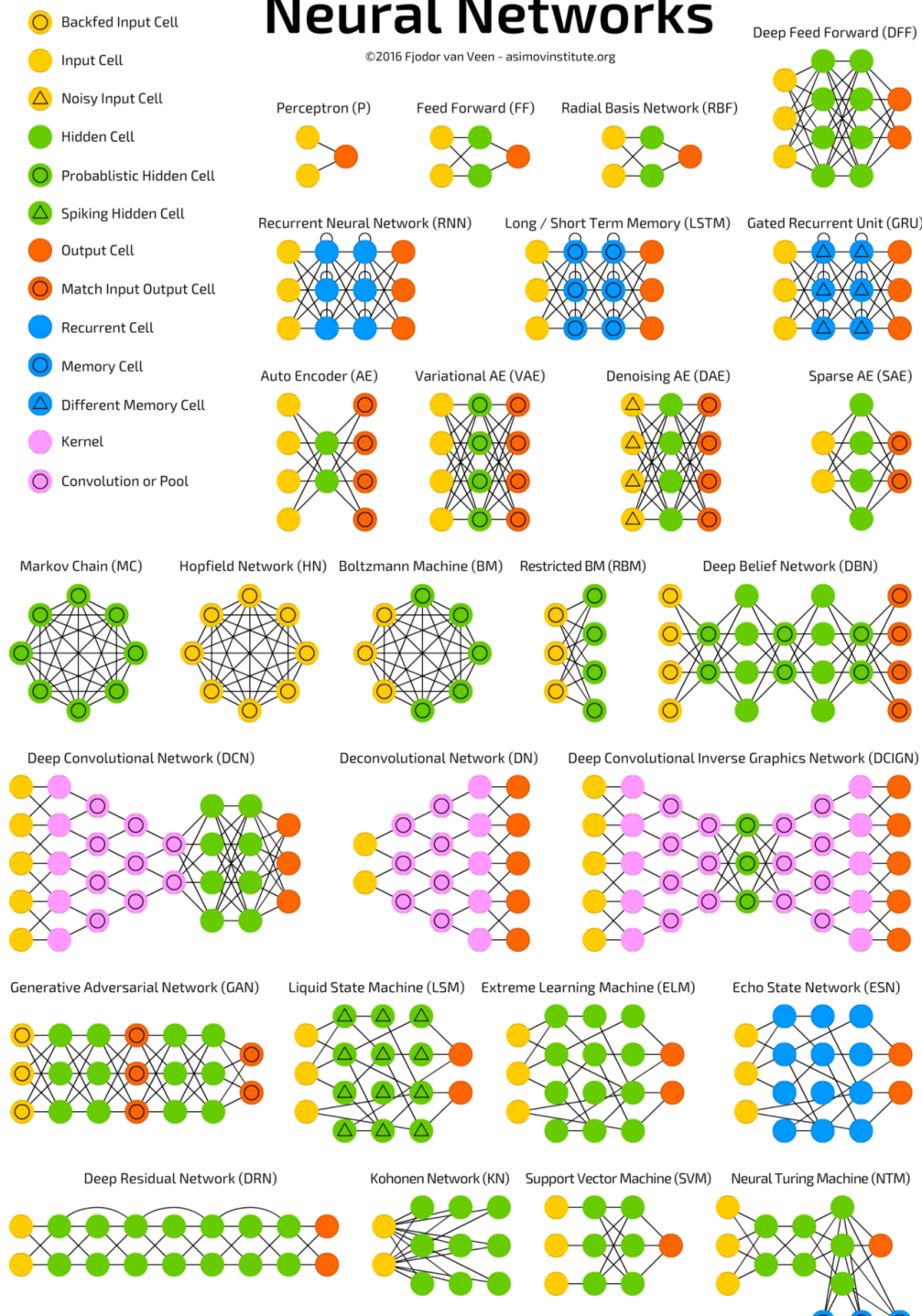
training data  
labels



criteria  
'pear'

## Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org



# motor



Model: relevante wereld kennis  
relevante criteria voor prestatie

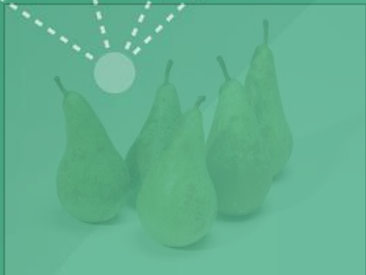


# Machine leren: de AI succes motor



'apple'

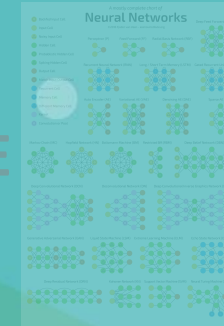
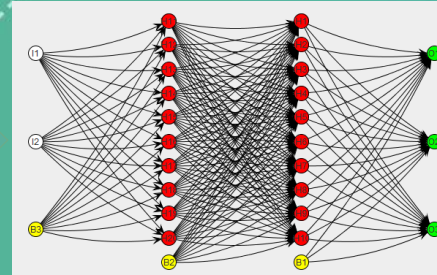
training data & labels



'pear'

criteria

LEARNING PHASE



APPLICATION PHASE

this is a ... ?





observed data

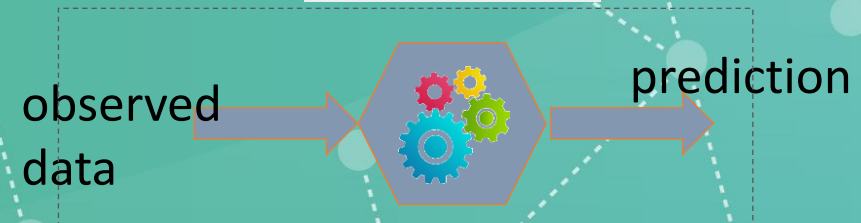
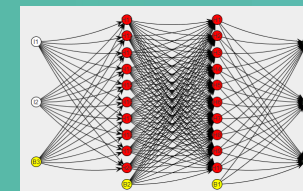






prediction

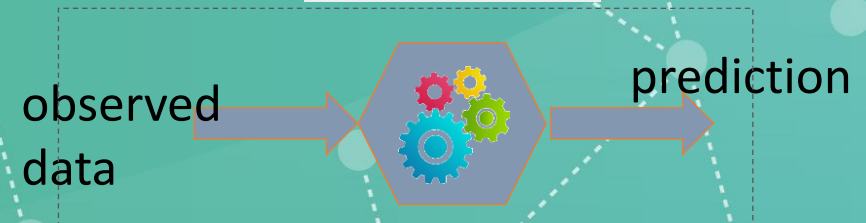
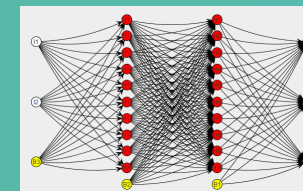
'pear'



		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		



		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		



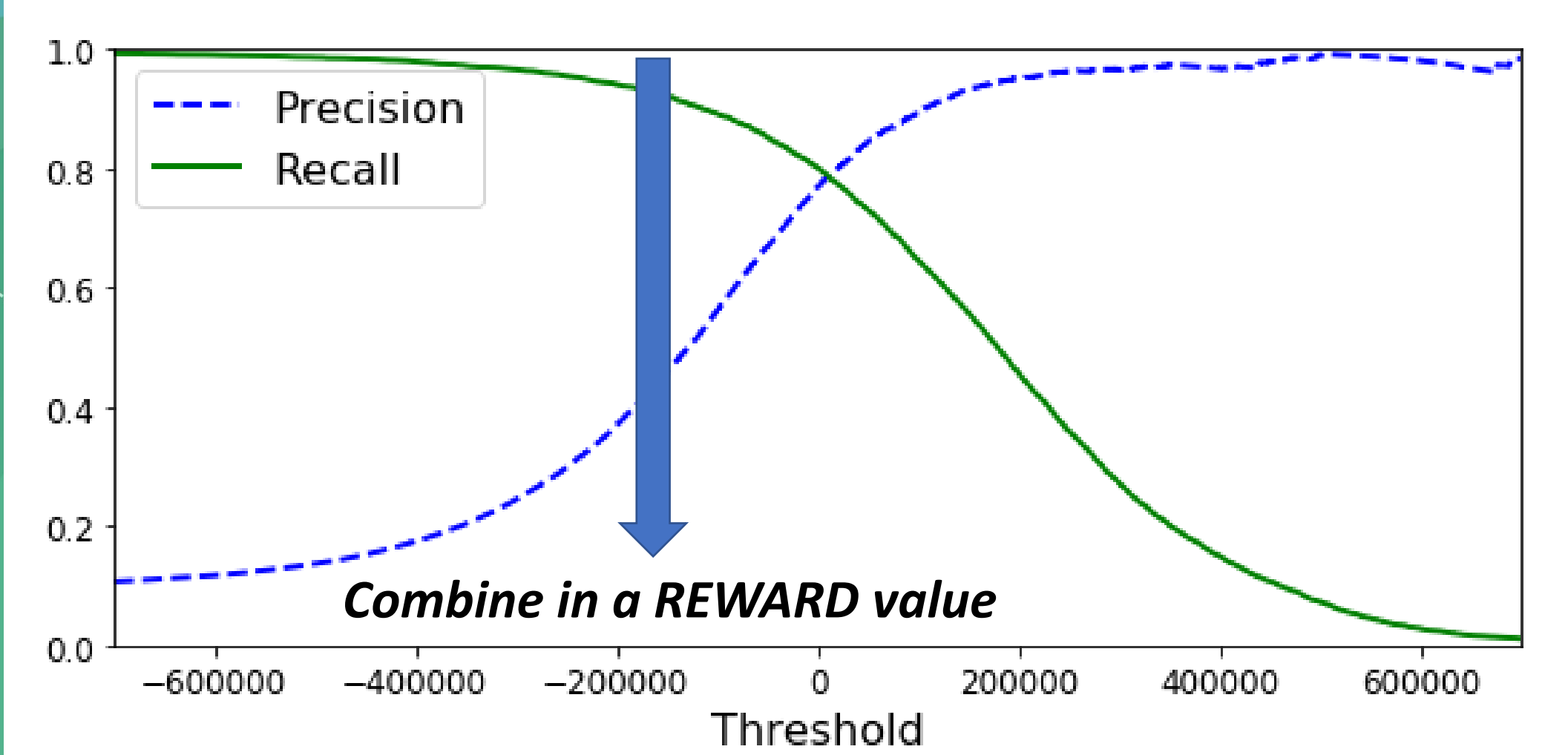
		ACTUAL (CONDITION)	
		Suitable	Not suitable
PREDICTED	Suitable	True positive	False positive
	Not suitable	False negative	True negative

*Note: A red dashed circle highlights the True positive and False negative cells. The word 'Precision' is written in blue to the right of the top-right cell, and 'Recall' is written in green below the bottom-left cell.*

$$Recall = TPR = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

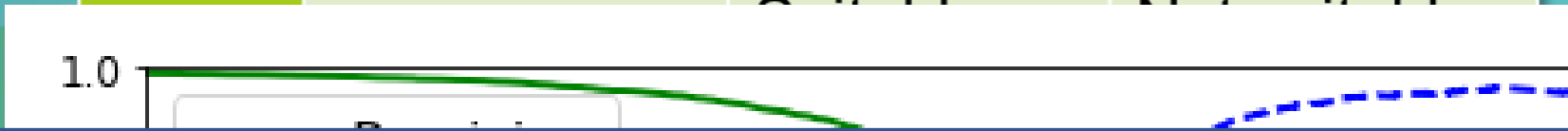
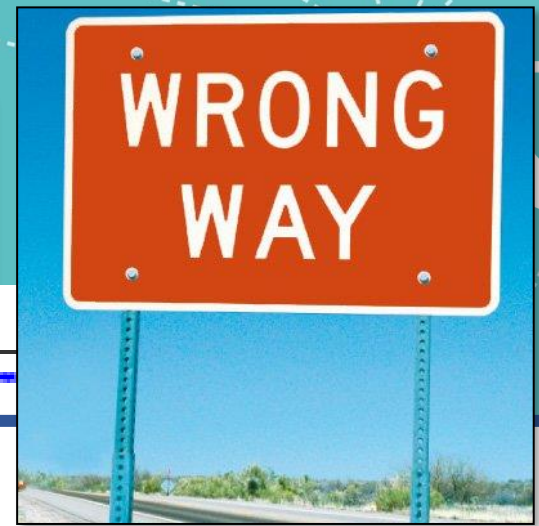
$$Precision = PPV = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

		ACTUAL (CONDITION)	
		Condition 1	Condition 2





		ACTUAL (CONDITION)



## Framing and Formalism trap

- Formulate goal of the system as an **optimization criterion** (AI-fication).
- What are you **optimizing** for? How much of the world does the data/model capture? (reward hacking and tunnel vision).

Threshold

MUST READ: [Google: Here's how we're toughening up Android security](#)

# Google: How do we build a cleaning robot that doesn't cheat or destroy things in its path?

The answer: It will be very difficult because once an AI robot figures out how to game the system, it won't be inclined to stop.

SCI-TECH SCIENCE TECHNOLOGY HEALTH AGRICULTURE ENVIRONMENT GADGETS

SCI-TECH > TECHNOLOGY

TECHNOLOGY

## Hackers can turn vacuum cleaners into spying devices, NUS researchers say



Sowmya Ramasubramanian

NOVEMBER 25, 2020 19:15  
UPDATED: NOVEMBER 25, 2020 19:18

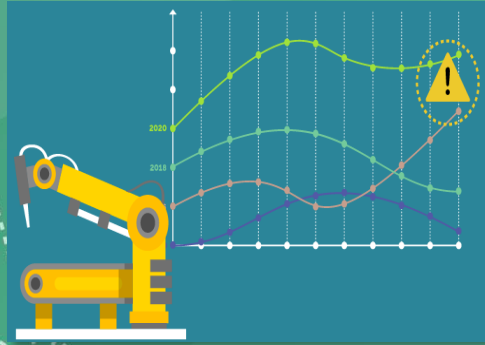
**COMPUTERWORLD** NETHERLANDS

NEWS ANALYSIS

**Google concerned about curious but destructive cleaning robots that hack reward systems**

Forget about AI triggering Skynet; researchers are concerned about curious cleaning robots with an el destructo twist and AI that hacks its reward system.

# Predictive maintenance



## Criterion

Estimate of repair costs

## Data

- \* vibration spectrum
- \* resource usage
- \* current signature

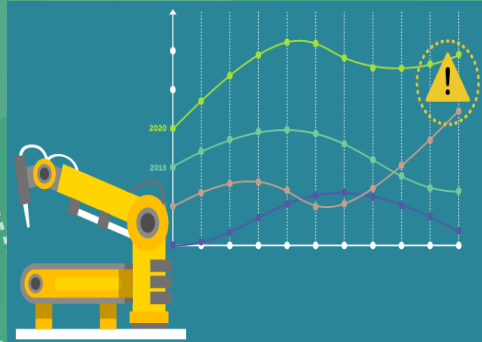
## Advantages

Maintenance planning

## Risks

Reduced value of expertise

## Predictive maintenance



## Life expectancy prediction



### Criterion

Estimate of repair costs

Risk score

### Data

- \* vibration spectrum
- \* resource usage
- \* current signature

- \* blood pressure
- \* creatine
- \* white blood cell count

### Advantages

Maintenance planning

Informed treatment decisions

### Risks

Reduced value of expertise

Root cause unclarity



## Accelerate recruiter productivity while providing exceptional experiences

Conversational AI automates monotonous recruiter tasks while providing candidates with the guidance and support they need throughout the entire hiring journey. In a self-guided and personalised experience, candidates interact with the virtual hiring assistant, allowing them to quickly find the job they are looking for, pre-screen for the role, schedule an interview, and receive automatic updates without any recruiter intervention. With candidate engagement taken care of, recruiters can focus on the most qualified candidates, quickly move them through the process, while spending more time on strategic priorities.

\* resource usage

\* creative

## Job suitability prediction



Avoid hiring unsuitable

- \* facial emotion
- \* voice timbre
- \* vocabulary

High volume selection  
Reduced self presentation

Criterion

Data

Advantages  
Risks

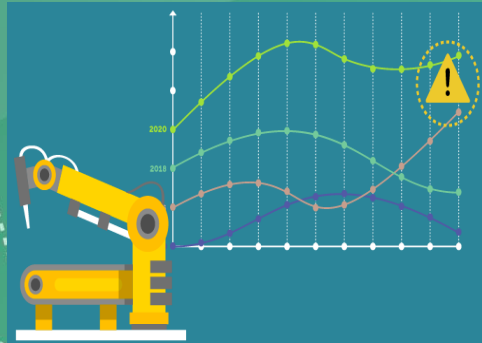
Work & Careers

+ Add to myFT

## 'Disease' of recruitment bias: is technology a cure or a cause?

Critics of AI platforms doubt whether they eliminate human interviewers' prejudices

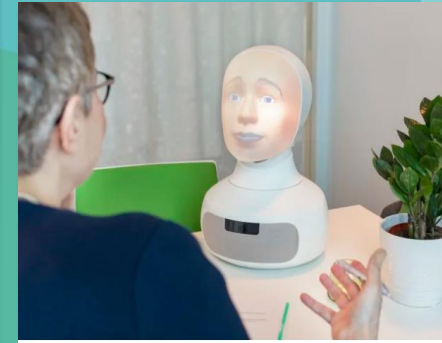
## Predictive maintenance



## Life expectancy prediction



## Job suitability prediction



## Sexual orientation prediction



Criterion

Estimate of repair costs

Risk score

Avoid hiring unsuitable

Reproduce human ability

# The invention of AI 'gaydar' could be the start of something much worse

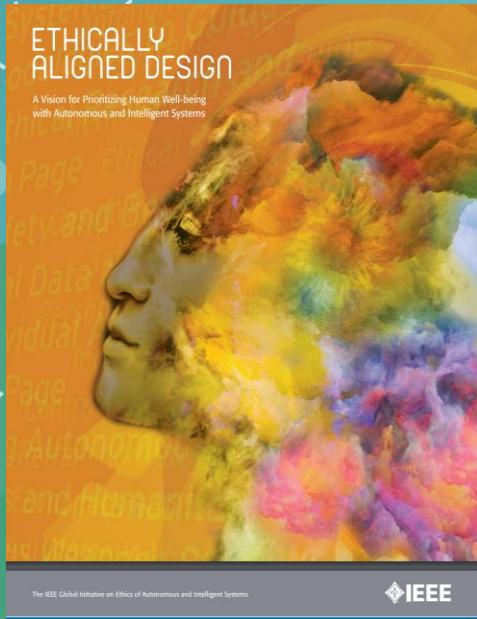
*Researchers claim they can spot gay people from a photo, but critics say we're revisiting pseudoscience*

By James Vincent | Sep 21, 2017, 1:24pm EDT

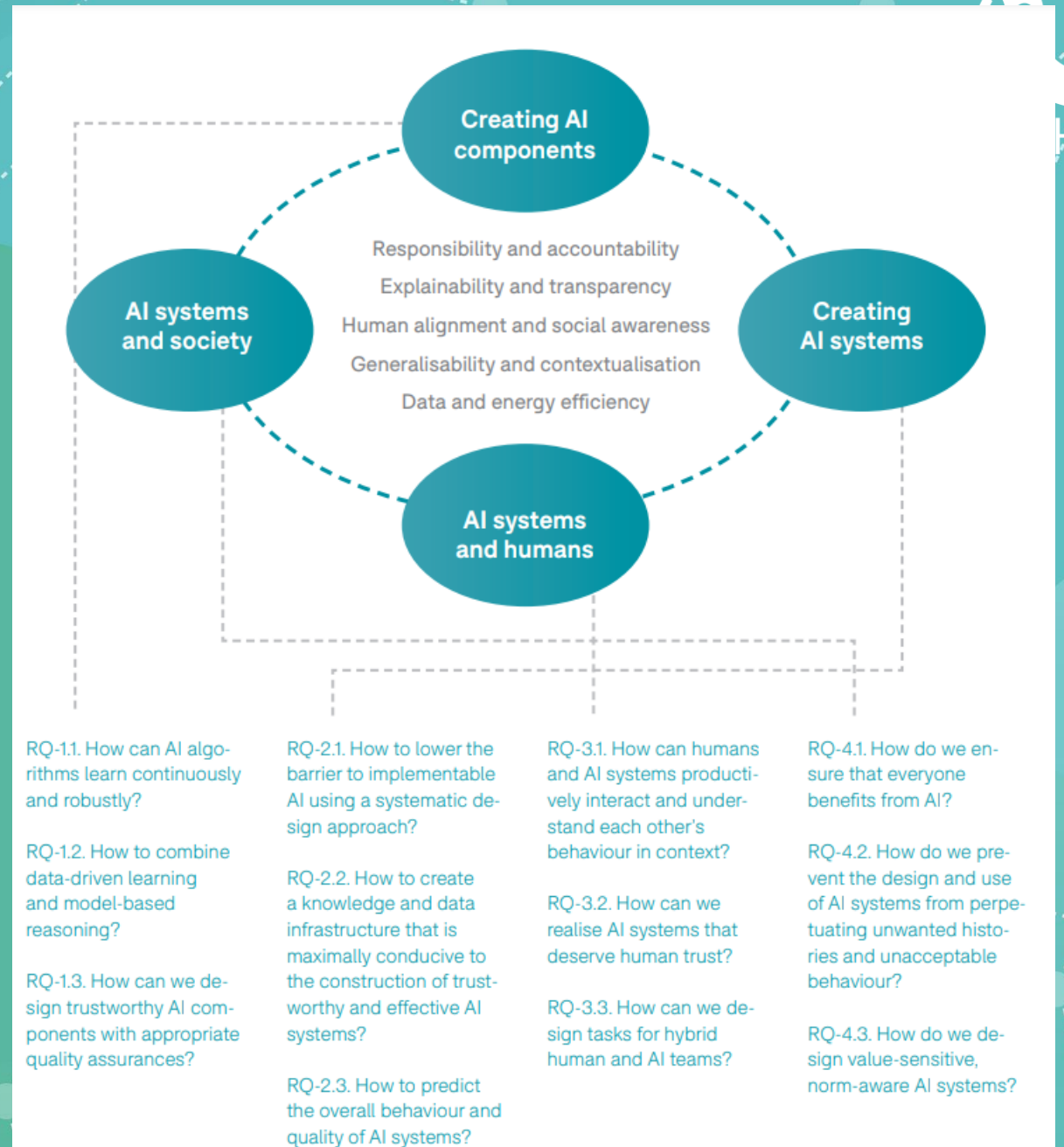
Illustrations by Alex Castro

- \* facial features
- \* morphology
- \* grooming

Security, marketing  
Stigmatization



## Human-centric, ethical & inclusive AI







- Autonomy
- Beneficence
- Justice



- Instrumental values relevant to autonomous (AI) systems





- Autonomy
- Beneficence
- Justice



```
81 self.cur_color = colors
82 self.mags_dt[i]
83
84 def on_key_press(self, symbol, modifiers):
85     """Delegate key press"""
86     if self.context_index == -1:
87         if symbol == key.UP and not self.active_index == 0:
88             self.menu_labels[self.active_index].color = [255, 255, 255, 255]
89             self.mags_dt = self.get_act_color_mag()
90         elif symbol == key.DOWN and not self.active_index == 3:
91             self.menu_labels[self.active_index].color = [255, 255, 255, 255]
92             self.active_index += 1
93             self.mags_dt = self.get_act_color_mag()
94         elif symbol == key.ENTER:
95             if self.active_index == 3:
96                 pygame.app.exit()
97             else:
98                 self.context_index = self.active_index
99         elif symbol == key.ESCAPE:
100             if self.context_index == -1:
101                 pygame.app.exit()
102             else:
103                 self.context_index = -1
104         elif self.context_index == 1:
105             if symbol == key.ESCAPE:
106                 self.context_index = -1
107             else:
108                 self.cur_game.on_key_press(symbol, modifiers)
109         else:
110             if symbol == key.ESCAPE:
111                 self.context_index = -1
```

- Instrumental values relevant to autonomous (AI) systems

*privacy*  
*transparency*  
*fairness*

...



AI ethical choices

GETTY

### Artificial intelligence

---

## In 2020, let's stop AI ethics-washing and actually do something

There's more talk of AI ethics than ever before. But talk is just that—it's not enough.

by **Karen Hao**

December 27, 2019

Last year, just as I was beginning to cover artificial intelligence, the AI world was getting a major wake-up call. There were some incredible advancements in AI research in 2018—from reinforcement learning to generative adversarial networks (GANs) to better natural-language understanding. But the year also saw several high-profile illustrations of the harm these systems can cause

- Autonomy
- Beneficence
- Justice



- Instrumental values relevant to autonomous (AI) systems

*privacy*  
*transparency*  
*fairness*

...



AI ethical choices

GETTY

### Artificial intelligence

## In 2020, let's stop AI ethics-washing and actually do something

There's more talk of AI ethics than ever before. But talk is just that—it's not enough.

by **Karen Hao**

December 27, 2019

Last year, just as I was beginning to cover artificial intelligence, the AI world was getting a major wake-up call. There were some incredible advancements in AI research in 2018—from reinforcement learning to generative adversarial networks (GANs) to better natural-language understanding. But the year also saw several high-profile illustrations of the harm these systems can cause

- Autonomy
- Beneficence
- Justice



- Instrumental values relevant to autonomous (AI) systems



- Current and possible technological capabilities



	Qualified person	Not-qualified person
Invited for interview	90	15
Not-invited for interview	10	85

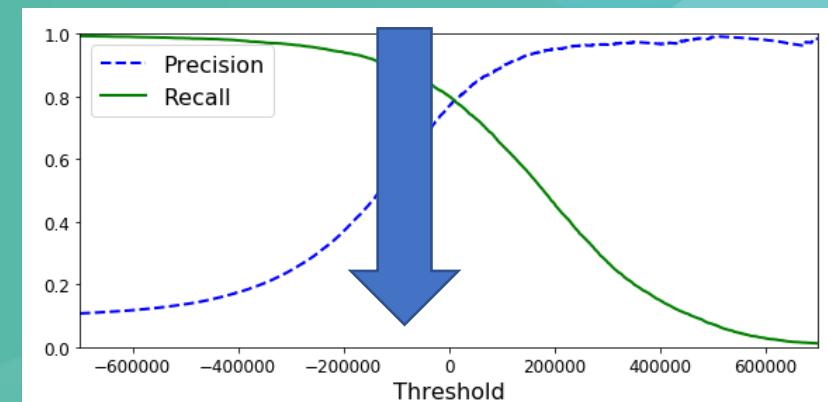




	Qualified person	Not-qualified person
Invited for interview	100	0
Not-invited for interview	0	100

	Qualified person	Not-qualified person
Invited for interview	90	15
Not-invited for interview	10	85

**REWARD = 0.87**



# ALL

	Qualified person	Not-qualified person
Invited for interview	90	15
Not-invited for interview	10	85

# SUBGROUP A

	Qualified person	Not-qualified person
Invited for interview	95	20
Not-invited for interview	5	80

# SUBGROUP B

	Qualified person	Not-qualified person
Invited for interview	75	10
Not-invited for interview	25	90

# ALL

	Qualified person	Not-qualified person
Invited for interview	90	15
Not-invited for interview	10	85

# SUBGROUP A

	Qualified person	Not-qualified person
Invited for interview	95	20
Not-invited for interview	5	80

# SUBGROUP B

	Qualified person	Not-qualified person
Invited for interview	75	10
Not-invited for interview	25	90



# ALL

	Qualified person	Not-qualified person
Invited for interview	90	15
Not-invited for interview	10	85

# SUBGROUP A

	Qualified person	Not-qualified person
Invited for interview	95	20
Not-invited for interview	5	80

+ (circled 20) + (circled 5)

# SUBGROUP B

	Qualified person	Not-qualified person
Invited for interview	75	10
Not-invited for interview	25	90

+ (circled 10) + (circled 25)

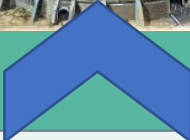
**Autonomous System for Good**  
Law and values: autonomy, beneficence, justice

Instrumental values: privacy, explainability, ...



**the new tower of Babel**

**many decisions are human**



Deep learning from data. Reasoning. Optimization, ...

**Autonomous System for Optimal Performance**  
Lower costs, better diagnosis, fewer malfunctioning, more mobility, ...

**Autonomous System for Good**  
Law and values: autonomy, beneficence, justice

**Rekenkamer: nauwelijks aandacht voor ethiek bij algoritmes overheid**

26 januari 2021 17:43

**the new tower of Babel**



**many decisions are human**

Deep learning from data. Reasoning. Optimization, ...

**Autonomous System for Optimal Performance**  
Lower costs, better diagnosis, fewer malfunctioning, more mobility, ...

# Meaningful Human Control

In order for humans to have, and be able to take control over the decisions of an autonomous system.



# Meaningful Human Control

In order for humans to have, and be able to take control over the decisions of an autonomous system.



human-AI system should be responsive to the human (moral) reasons relevant in the circumstances.

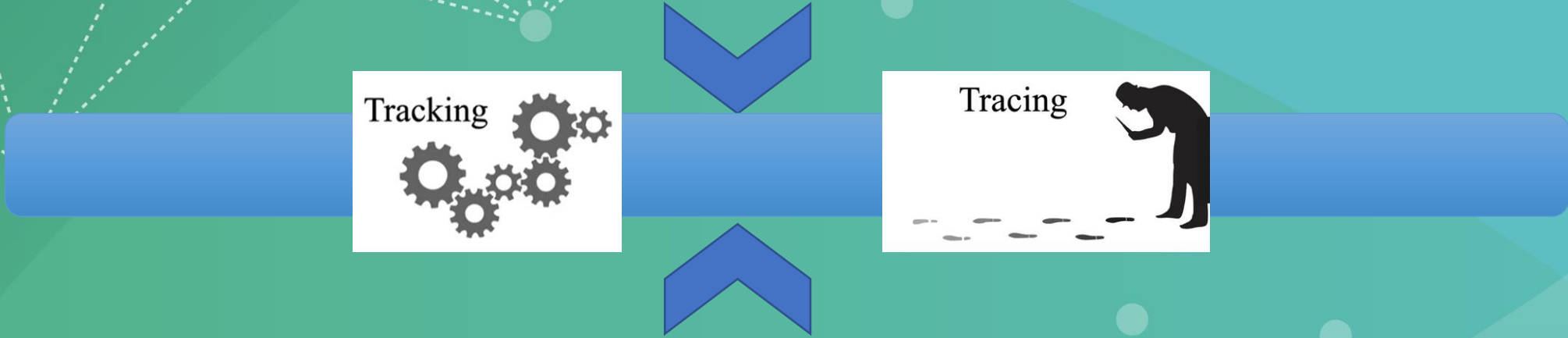


human-AI system behavior, capabilities, and possible effects in the world should be traceable to a proper moral understanding on the part of at least one relevant human agent who designs or interacts with the system.

# Autonomous System for Good

Law and values: autonomy, beneficence, justice

Instrumental values: privacy, explainability, ...



Deep learning from data. Reasoning. Optimization, ...

# Autonomous System for Optimal Performance

Lower costs, better diagnosis, fewer malfunctioning, more mobility, ...

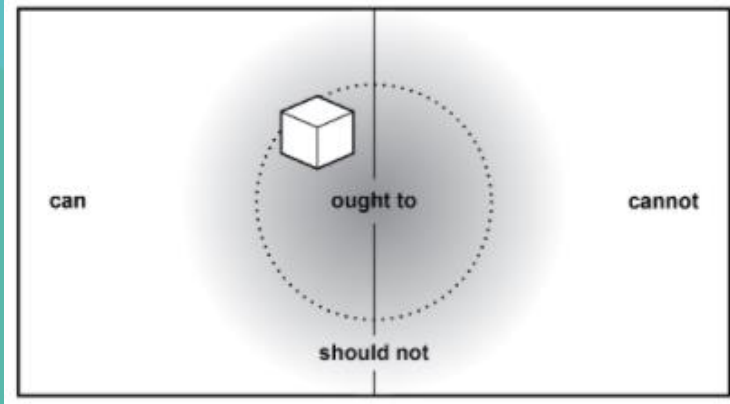
**Autonomous System for Good**  
Law and values: autonomy, beneficence, justice

Instrumental values: privacy, explainability, ...



**Responsibilities**

*Methodology → Moral Operational Design Domain*



Deep learning from data. Reasoning. Optimization, ...

**Autonomous System for Optimal Performance**  
Lower costs, better diagnosis, fewer malfunctioning, more mobility, ...

		Qualified person	Not-qualified person
<b>Inside M-ODD</b>	Invited for interview	75	10
<b>Outside M-ODD</b>	Human action	20	25
<b>Inside M-ODD</b>	Not-invited for interview	5	65

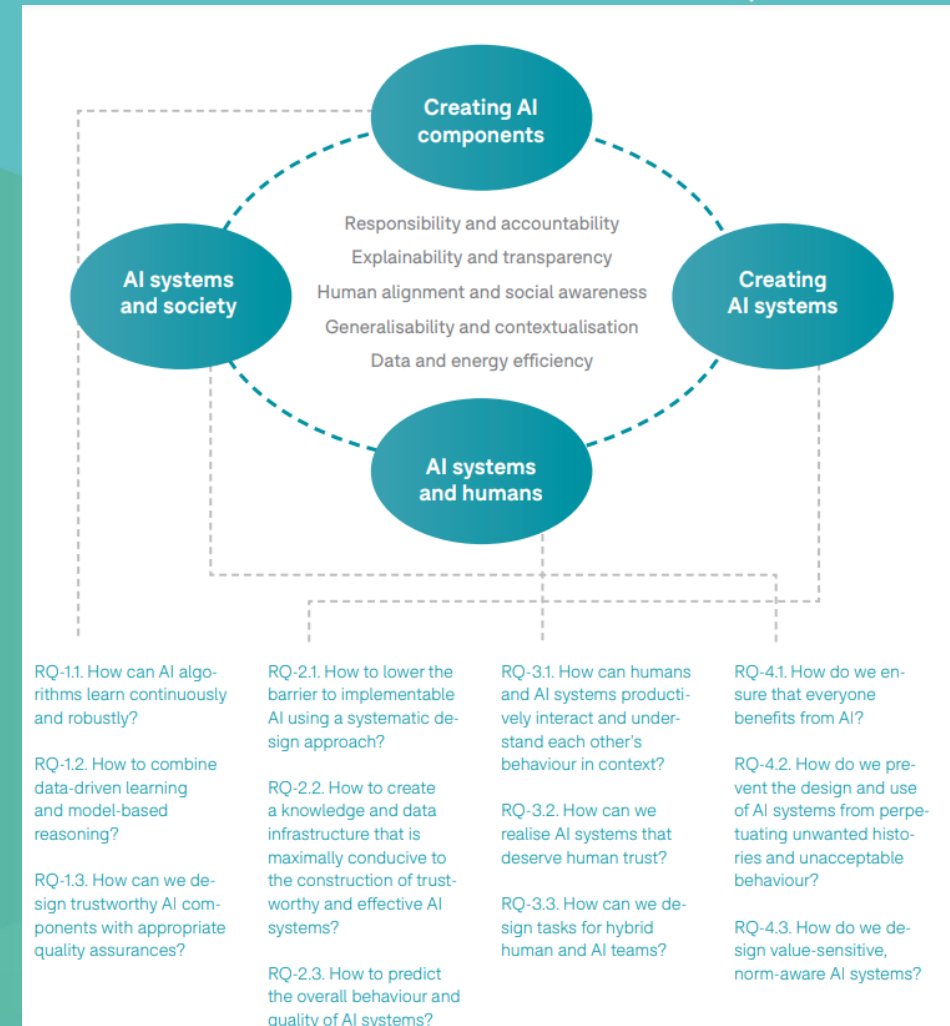
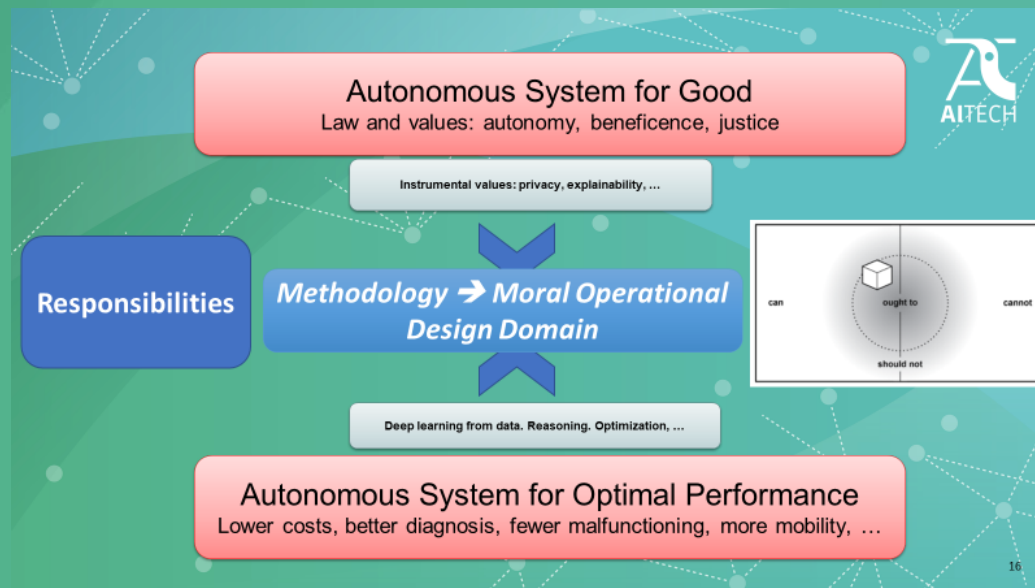


# Properties to Realize

- Alignment of ability, authority and responsibility.
- Adequate and compatible mental models.
- Actions of AI agents are explicitly linked to human decisions.
  
- And so:
  - Computational optimization is not always the solution.
  - Design AI systems for minimal invasive effect (use only when and where needed).
  - AI systems should eventually be aware of their own limitations.

# Needed for Operational Control

- Data collection and analysis.
- Cognitive models and interaction.
- Symbolic and statistical (moral) reasoning.
- Fairness, transparency, explainability.





Visit us at <https://www.tudelft.nl/aitech/>



The image shows a screenshot of the TU Delft website. At the top is a blue navigation bar with the TU Delft logo on the left and menu items: AiTech, Projects, Publications, News, Our team, Work with us, and Agora. A search icon is on the right. Below the navigation bar is a large banner with a diverse crowd of stylized human faces. Overlaid on the banner is the text: "AiTech paper 'Designing for Human Rights in AI' published in Big Data & Society". A "READ MORE" button is visible on the left side of the banner. Below the banner is a dark blue bar with the text: "AiTech is TU Delft's multidisciplinary research program on awareness, concepts, and design & engineering of autonomous technology under meaningful human control".

### Why meaningful human control?

Today's engineers create systems that are ever more equipped with artificial intelligent technologies. Autonomous behavior of cars, robots, and decision support algorithms is becoming a reality. Our vision is that scientists should not only research the technology that makes

### Our 'how to' approach

Meaningful human control is particularly important in cases of failures or conflicts with the normative foundations of society, social conventions, and human acceptability. We believe these challenges demand a multidisciplinary effort, bringing together researchers across



26/01/2021

## Robot journalists make inroads in newsrooms, but not without human colleagues

[Culture](#), [News](#), [Societal acceptance and inclusion](#)

[→ Read more](#)

### Latest news

15/01/2021

**NL AIC works on international cooperation aimed at economic opportunities**

[News](#)

23/12/2020

**AI makes driving a car safer, but steering it yourself still feels better**

[Mobility, Transportation and Logistics](#), [Societal acceptance and inclusion](#)

11/12/2020

**How robots make agriculture more sustainable**

[Agriculture and Nutrition](#), [News](#), [Societal acceptance and inclusion](#)

[→ Latest news](#)

**ELSA Labs** zijn participatieve innovatie omgevingen waarin AI en data technologie en toepassingen in kaart gebracht en gevalideerd worden op een manier die borgt dat alle relevante groepen betrokkenen de toepassingen als zinvol, haalbaar, verantwoord en wenselijk ervaren



FIN