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

rtl nieuws

26 januari 2021 17:43
Aangepast: 26 januari 2021 18:07

Er gaat van alles mis met het gebruik van algor overheid, concludeert de Algemene Rekenkam rapport. Op tal van ministeries wordt bijvoorbe besteed aan de ethische aspecten van algoritn discriminatie.

Op slechts drie van alle twaalf ministeries staat prioriteitenlijst als er wordt nagedacht over een gesprekken die de Rekenkamer voerde met mec ministeries. Welke ministeries dat zijn, wil de Rel onafhankelijke controleur van de overheid, over was niet het doel van het onderzoek, legt een w

Het doel was vooral om te kijken hoe het gebrui overheid is geregeld. En dat is niet best, blijkt. O ministeries vaak niet eens welke algoritmes er g

Wat is een algoritme?

Een algoritme is een verzameling regels die ee op grote schaal taken te verrichten die menser Zo kan een algoritme bijvoorbeeld gebruikt wo sporen, door allerlei gegevens door te spitten wijzen.

Geen overzicht

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

Rekenkamer: ministers weten niet welke algoritmes ze gebruiken

Jan Fred van Wijnen 26 jan

fd.

Discriminatie op de loer

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

Algoritmes zijn programma's die zelfstandig data verwerken. Daardoo 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 febru 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 essentiele 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**.

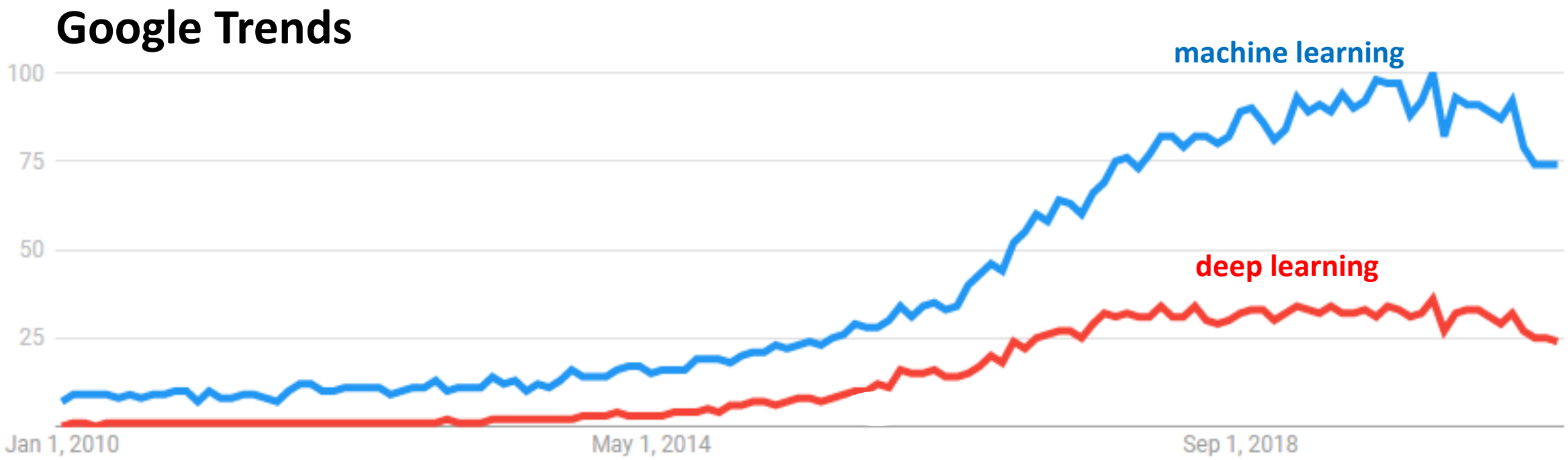
Voordelen. Risicos. Verantwoordelijkheid

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

Hoe slim is het (AI) algoritme?

Kan het (AI) algoritme deze verantwoordelijkheden dragen?

Machine leren: de AI succes motor

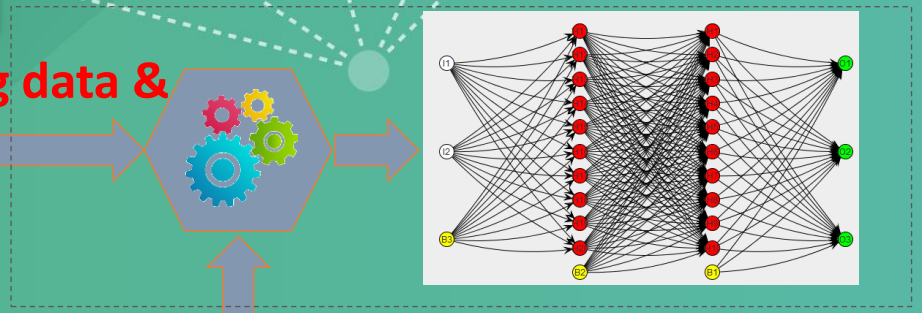


Machine leren: de AI succes motor

‘apple’

LEARNING PHASE

training data &
labels



criterion

‘pear’

Model:

1. De relevanten wereld kennis

2. Relevante criteria voor prestatie

Machine le motor

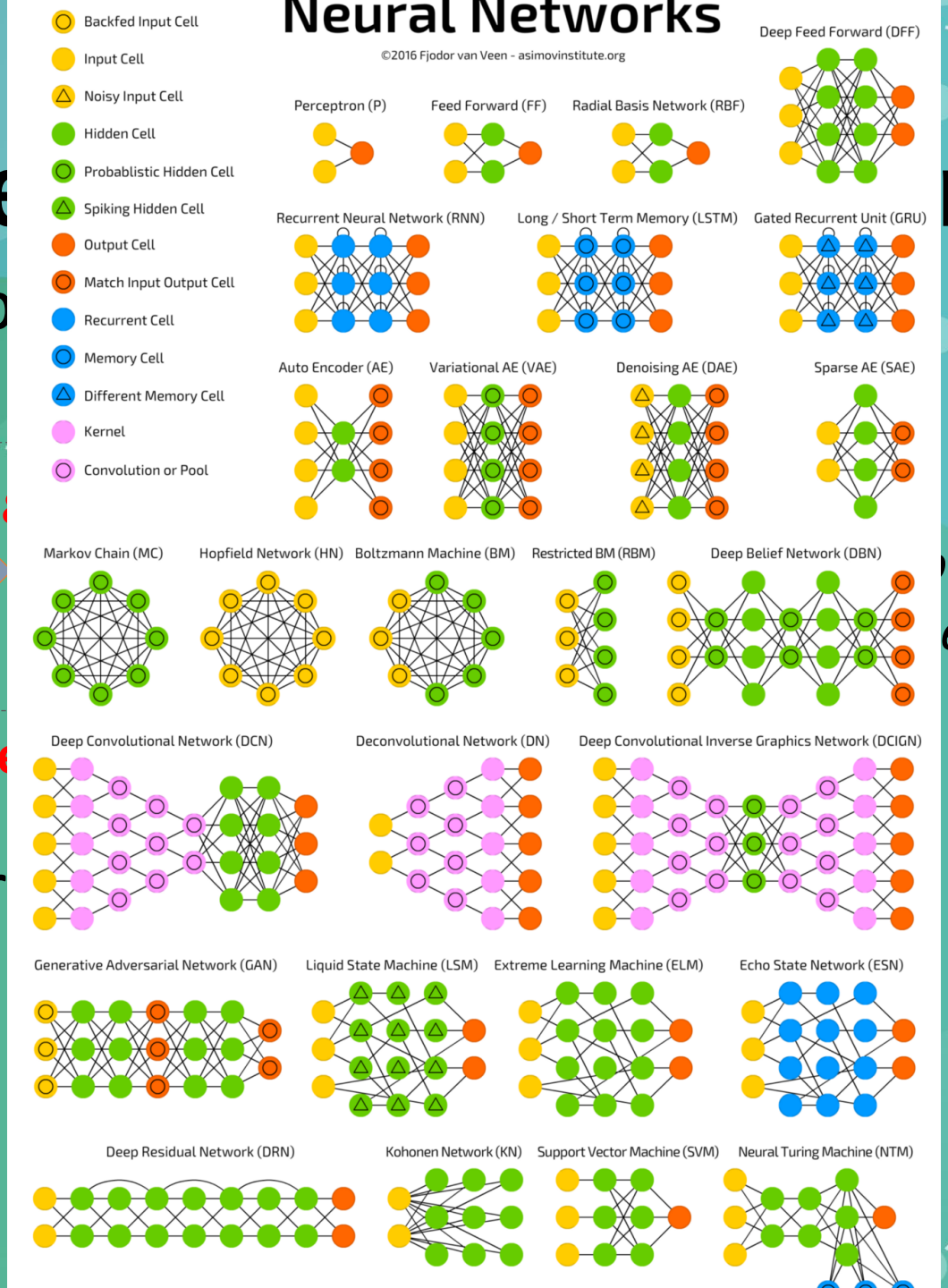
'app

training data
labels

crite

'pear

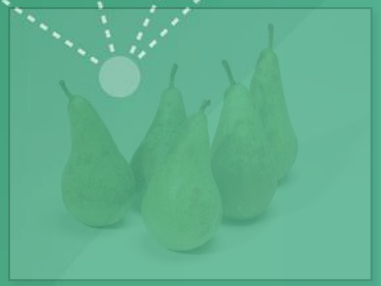
del: relevante wereld kennis
evante criteria voor prestatie



Machine leren: de AI succes motor

‘apple’

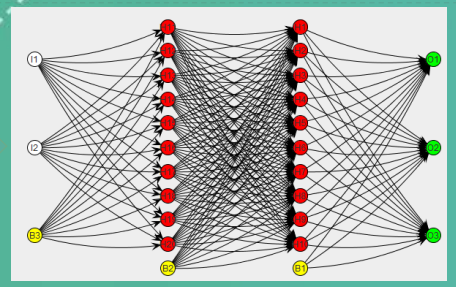
training data & labels

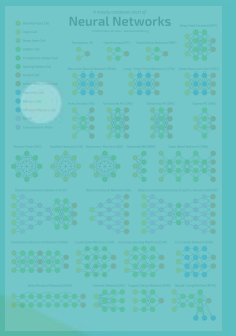


criterion

‘pear’

LEARNING PHASE





APPLICATION PHASE

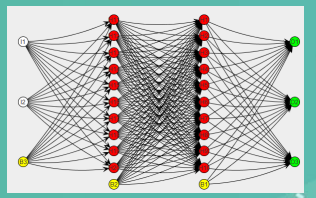
this is a … ?

observed data

prediction

‘pear’

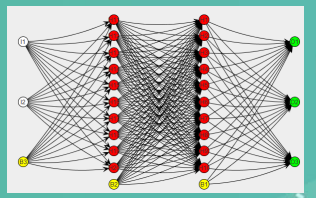
		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		

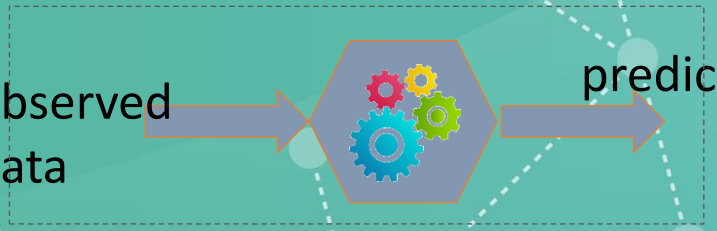


observed data

prediction

		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		





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

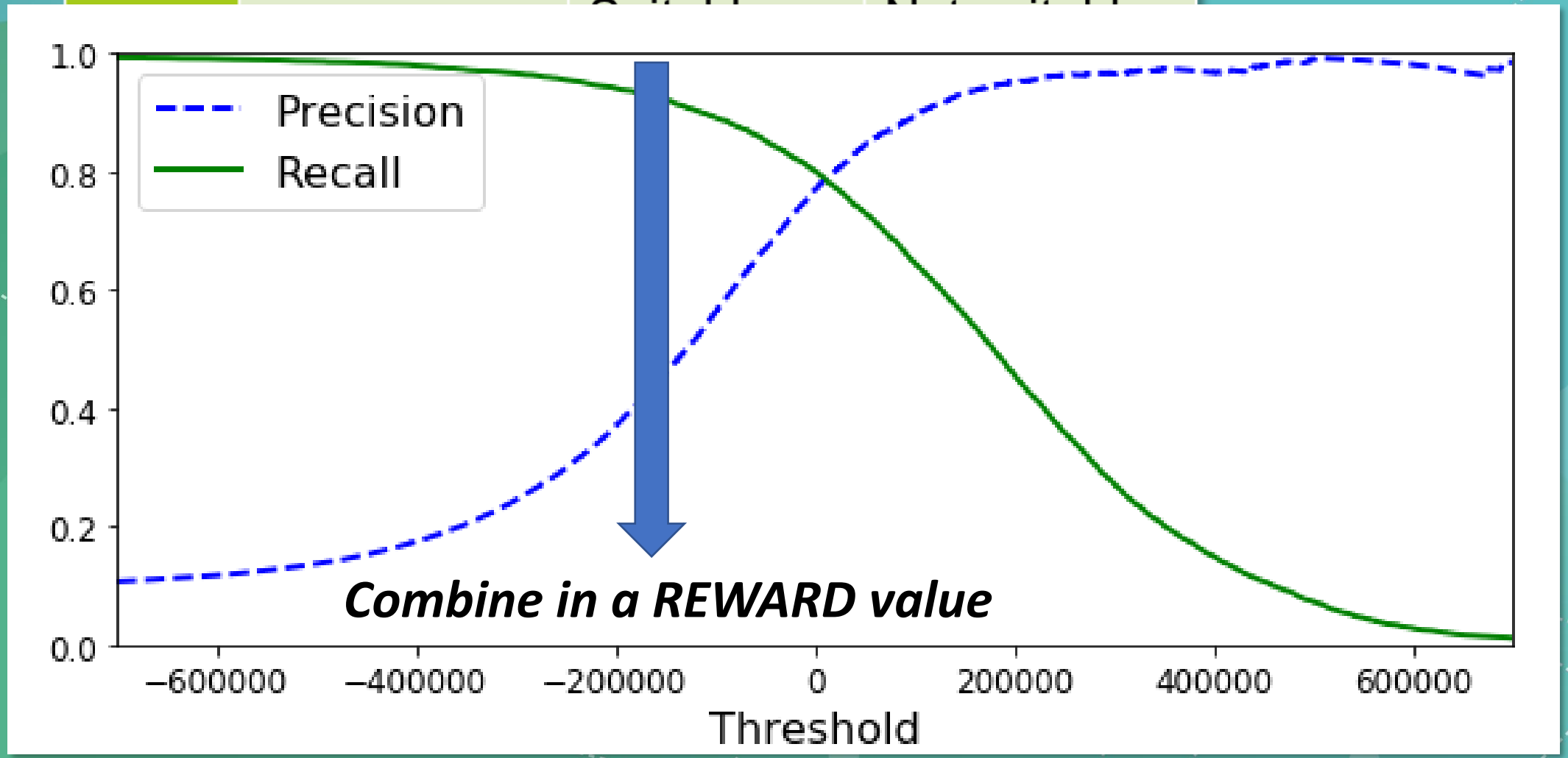
Precision

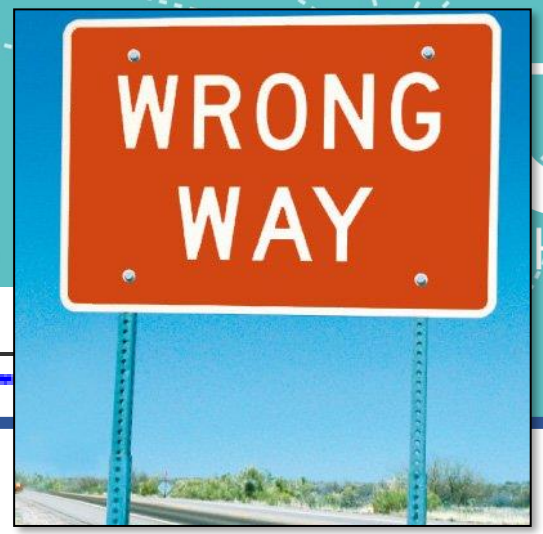
Recall

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

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

		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).

ZDNet

CENTRAL EUROPE MIDDLE EAST SCANDINAVIA AFRICA UK ITALY SPAIN MORE NEWSLETTERS ALL WRITE

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

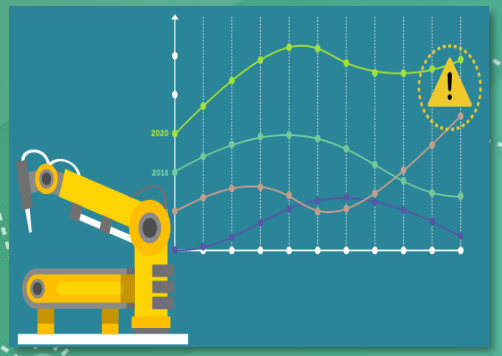
COMPUTERWORLD NETHERLANDS

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



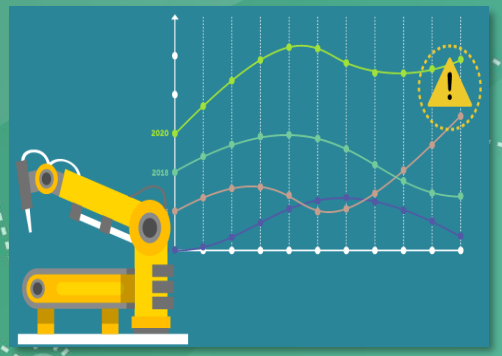
Criterion Estimate of repair costs

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

HireVue

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.

Criterion

Data

* resource usage

* creatine

Advar
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

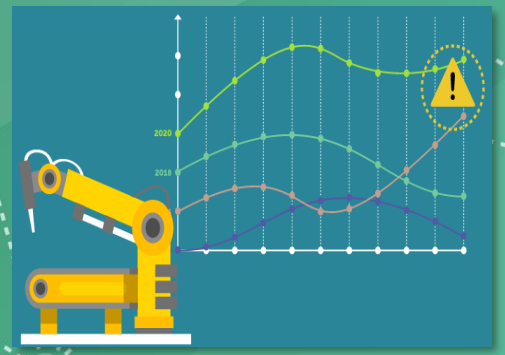
Job suitability prediction

Avoid hiring unsuitable

* facial emotion
* voice timbre
* vocabulary

High volume selection
Reduced self presentation

AITECH

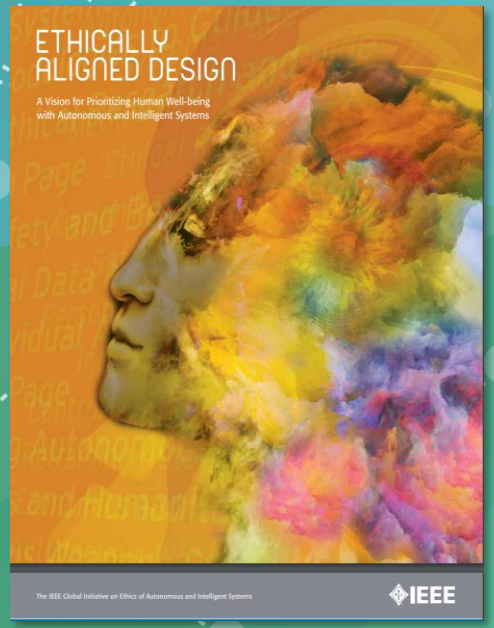
	Predictive maintenance	Life expectancy prediction	Job suitability prediction	Sexual orientation prediction
				
Criterion	Estimate of repair costs	Risk score	Avoid hiring unsuitable	Reproduce human ability
				* facial features * morphology * grooming
				Security, marketing Stigmatization

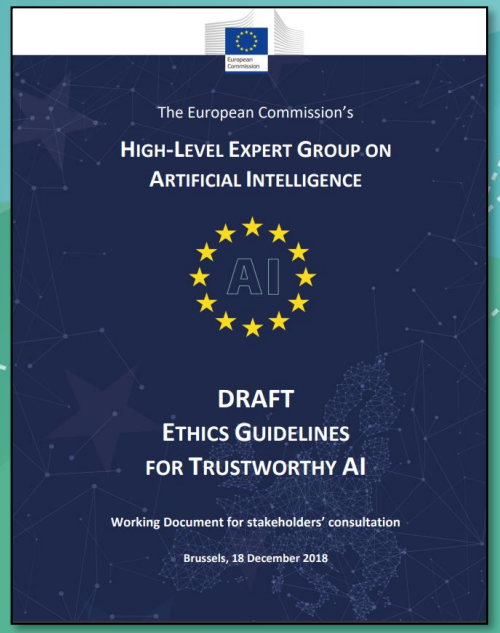
The invention of AI 'gaydar' could be the start of something much worse

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

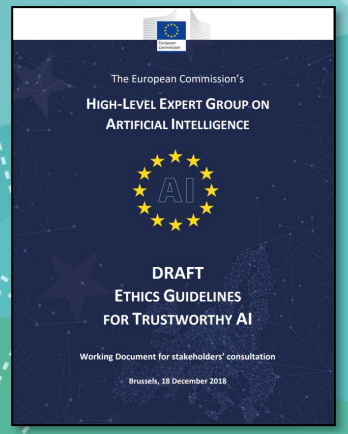
By James Vincent | Sep 21, 2017, 1:24pm EDT

Illustrations by Alex Castro



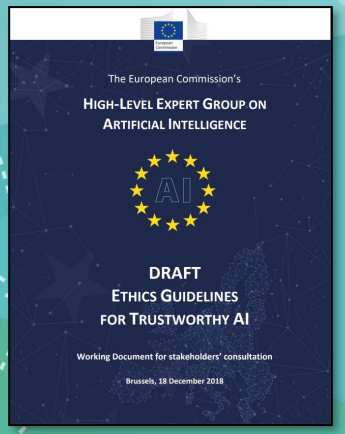


Human-centric, ethical & inclusive AI



- Autonomy
- Beneficience
- Justice

- Instrumental values relevant to autonomous (AI) systems



- Autonomy
- Beneficience
- 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

- Autonomy
- Beneficience
- 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

- Autonomy
- Beneficience
- Justice

- Instrumental values relevant to autonomous (AI) systems

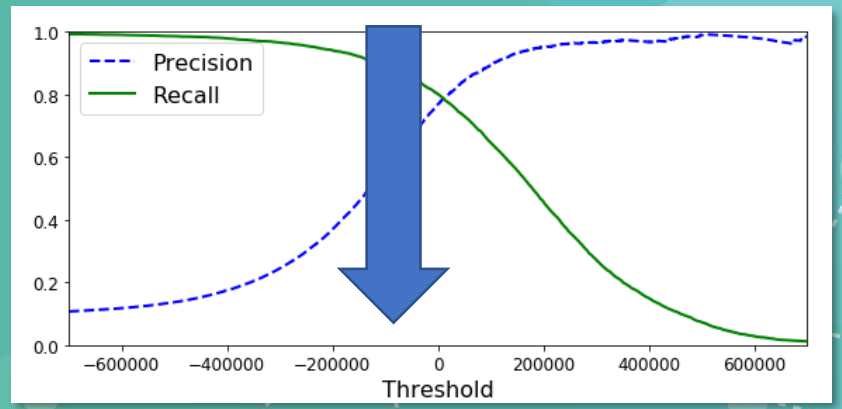
- Current and possible technological capabilities

	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

SUBGROUP B

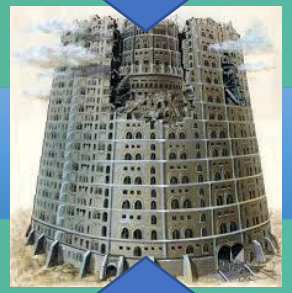
	Qualified person	Not-qualified person
Invited for interview	75	10
Not-invited for interview	25	90

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, …

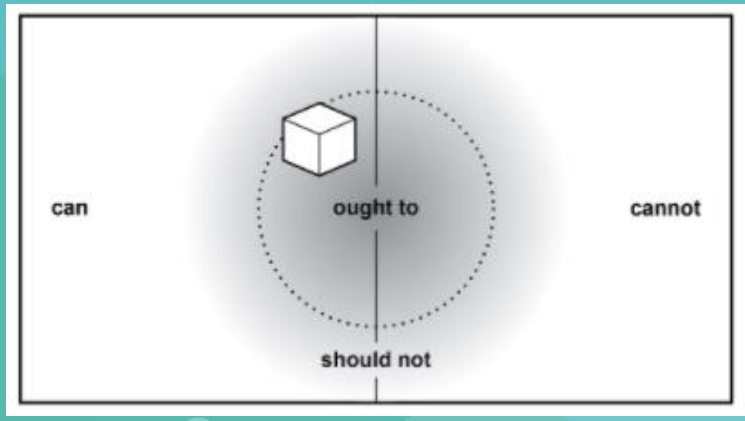
Autonomous System for Good

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
Inside M-ODD	Invited for interview	75	10
Outside M-ODD	Human action	20	25
Inside M-ODD	Not-invited for interview	5	65

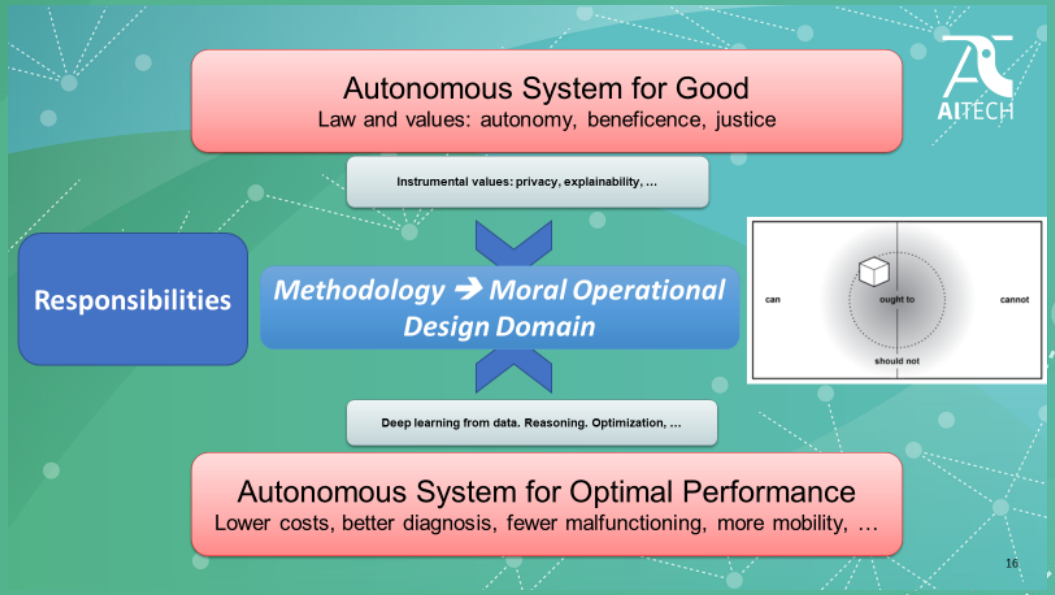
Properties to Realize

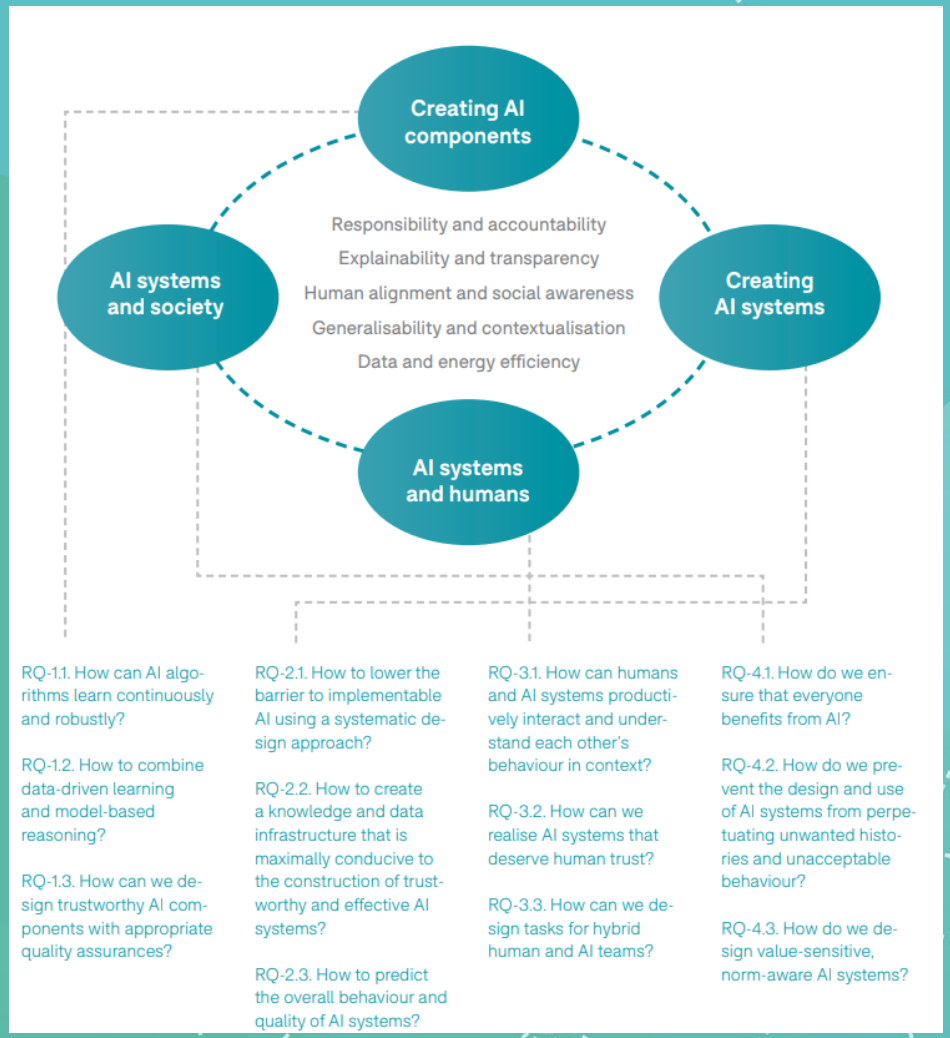
- Alignment of ability, authority and responsibility.
- Adequate and compatible mental models.
- Actions of AI agents are explicitly linked to human decisions.

- 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.

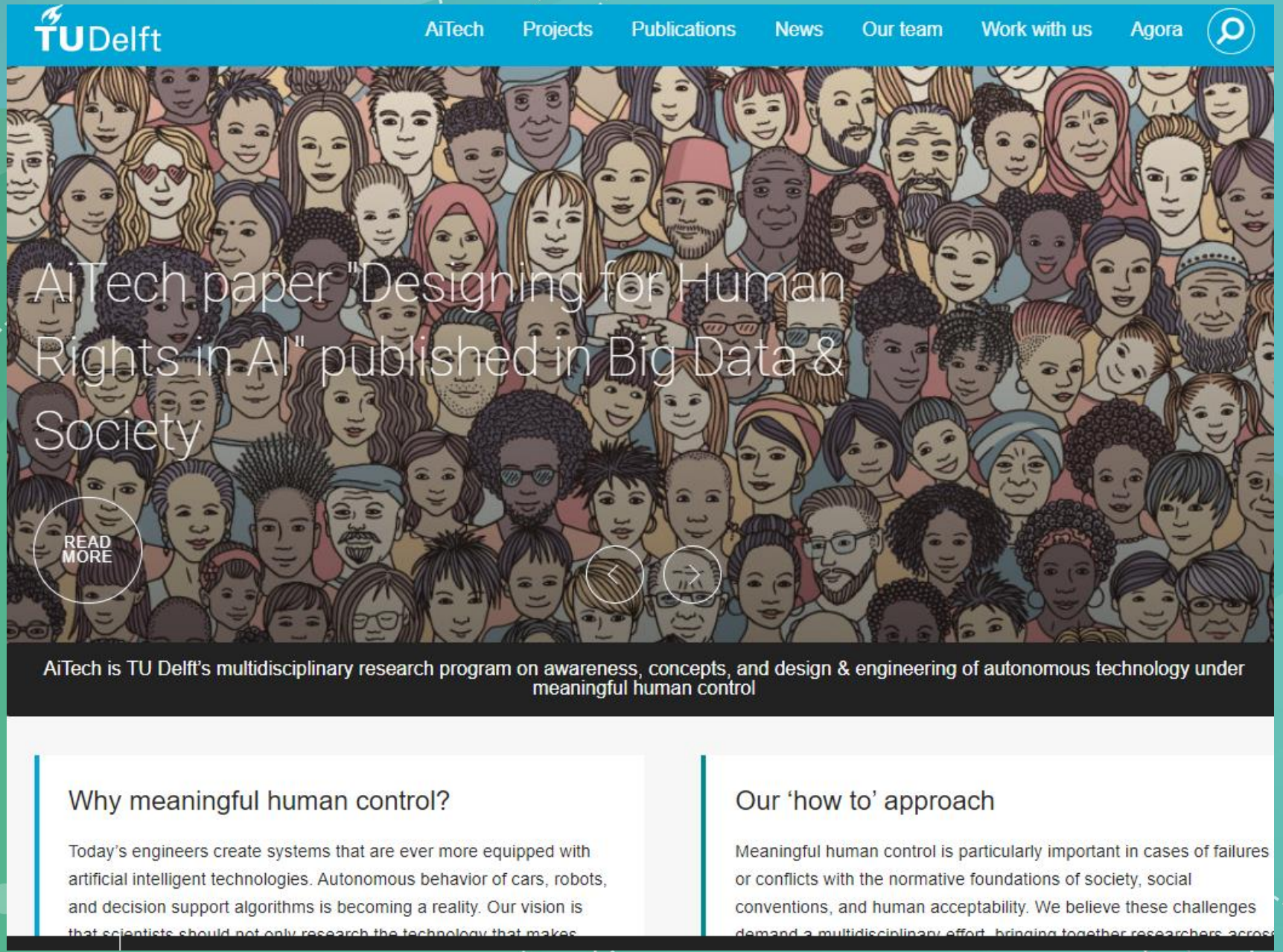
Needed for Operational Control

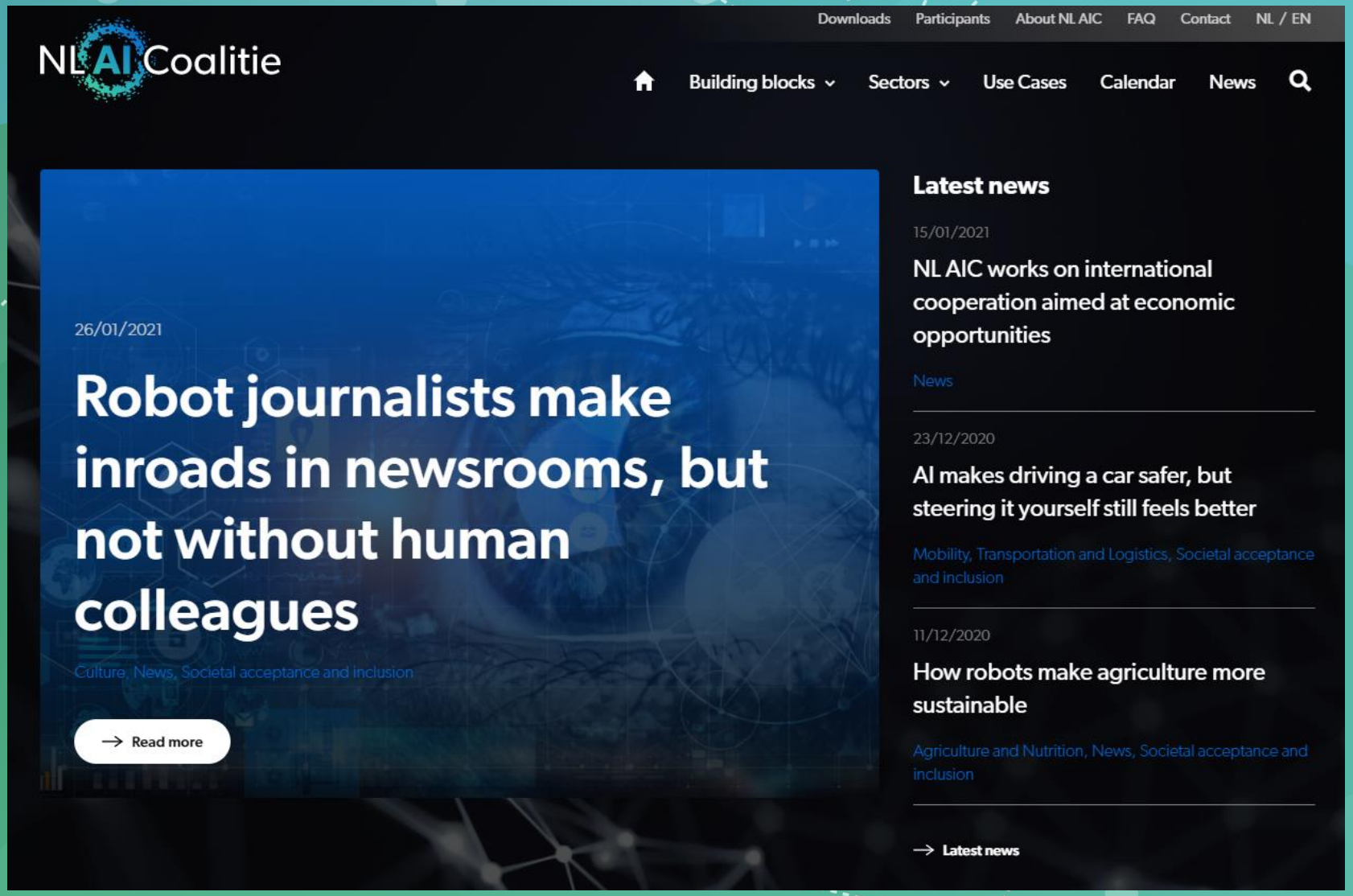
- Data collection and analysis.
- Cognitive models and interaction.
- Symbolic and statistical (moral) reasoning.
- Fairness, transparency, explainability.





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





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