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 algoritmes door o
overheid, concludeert de Algemene Rekenkamer in een krit
rapport. Op tal van ministeries wordt bijvoorbeeld geen aa
besteed aan
discriminatie

Op slechts dr
prioriteitenlijs
gesprekken d
ministeries. W
onafhankelijk
was niet het c

Het doel was
overheid is ge
ministeries vo

Wat is ee

Een algoritm
op grote sch
Zo kan een c
sporen, doo
wijzen.

Geen ove

Pas na het be
veel ministeri
welke algoritmes er eigenlijk al gebruikt worden.

Rekenkamer: ministers wete
welke algoritmes ze gebruike

Jan Fred van Wijnen 26 Jan

fd.

EUROPA

fd. Het Financieele Dagblad
Donderdag 22 april 2021

EU legt kunstmatige intelligentie gedragsregels op

Mathijs Schiffers
Brussel

Een mijlpaal, noemt Eurocommissaris Margrethe Vestager het. De Deense presenteerde gisteren het langverwachte Brusselse voorstel voor het reguleren van kunstmatige intelligentie.

Het is 's werelds eerste poging om de ongewenste gevolgen van de zich in volle vaart ontluikende technologie aan banden te leggen. En Brussel hoopt dat goed voorbeeld zal doen volgen.

'Vertrouwen is noodzakelijk bij kunstmatige intelligentie en niet iets dat leuk is om erbij te hebben', zei Vestager. Ze bestrijdt dat Europa het risico loopt terrein te verliezen bij de ontwikkeling van toepassingen gebaseerd op kunstmatige intelligentie door als enige beperkingen op te werpen. Ze liet juist weten bang te zijn dat mensen de beloftevolle technologie links laten liggen als er geen paal en perk gesteld wordt aan de negatieve aspecten ervan.

Kunstmatige intelligentie verwijst naar het leervermogen van computers. Die kunnen snel grote hoeveelheden data verwerken, wat handig kan zijn bij het bepalen van iemands kredietwaardigheid of geschiktheid voor een baan. Maar kwaadwillenden kunnen de systemen zo inrichten dat bij de besluiten bijvoorbeeld sprake is van etnisch profileren en andere methodes die op gespannen voet staan met de fundamentele Europese waarden, zoals het recht op privacy.

Het gros van de toepassingen is volgens de Europese Commissie risicoloos. Daar zijn geen aanvullende eisen voor nodig of slechts beperkte, zoals bij het gebruik van chatbots. In dat geval wil de Commissie dat de gebruiker weet dat hij met een computer praat.

Maar voor sommige toepassingen is meer vereist. Die kennen een 'hoog risico' van misbruik, zoals computersystemen die worden ingezet door overheden die willen vaststellen of iemand voor een uitkering in aanmerking komt. Voordat zo'n systeem mag worden gebruikt, moet een test onder meer uitwijzen of het transparant en nauwkeurig is en of menselijk toezicht mogelijk is.

> Systemen met een hoog risico op misbruik moeten op transparantie worden getest

Daarnaast is er nog een aantal toepassingen dat volgens de Commissie ronduit verboden moet worden, zoals systemen die het gedachtegoed van iemand proberen te manipuleren of de kwetsbaarheid van kinderen exploiteren. Bedrijven die zich niet aan de regels houden, riskeren boetes tot €30 mln of 6% van de mondiale omzet, al naar gelang wat hoger is.

Er is bijval en kritiek voor het plan. 'Een evenwichtig startpunt', zegt Lawrence Kerknawi van public-affairsbureau Political Intelligence in Brussel, dat opkomt voor de belangen van de techsector. 'Het is goed dat Brussel een streep in het zand zet en zegt: dit vinden we onaanvaardbaar', zegt Bart Schermer van het Nederlandse public-affairsbureau Considerati.

Schermer maakt zich wel wat zorgen over 'de praktische uitvoerbaarheid' van de plannen. 'Organisaties zullen een flinke professionalisering moeten doorvoeren en het tekort aan datawetenschappers zal er niet minder op worden, nu de toezichthouders ze ook nodig zullen hebben.'

Edri, dat zich inzet voor rechten en vrijheden op internet, is vooral teleurgesteld in het feit dat de Commissie 'niet elke vorm van biometrische massasurveillance' heeft verboden.

PvdA-Europarlementariër Paul Tang vindt ook dat het voorstel niet ver genoeg gaat. 'Kunstmatige intelligentie is een belangrijke motor voor innovatie, maar de motor kan alleen draaien als de passagiers beschermd zijn tegen ongelukken.'

De wetgeving wordt naar verwachting in 2023-24 van kracht.

Uber's self-driving operator charged over fatal crash

BBC NEWS

16 September 2020

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

icycle across

een streaming

hat there was

ng car, and

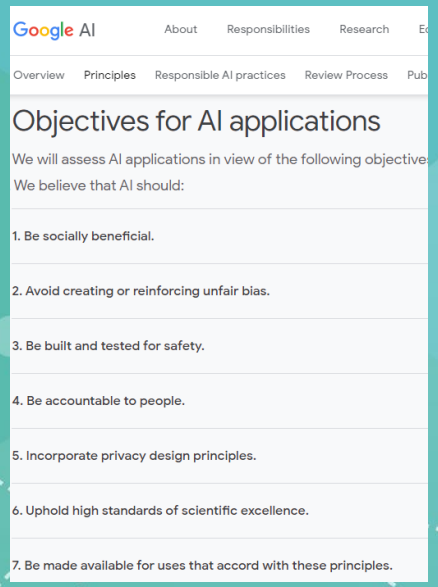
tation Safety
the crash.

over control of

ing down, away
h, while the car

Het AI Debat

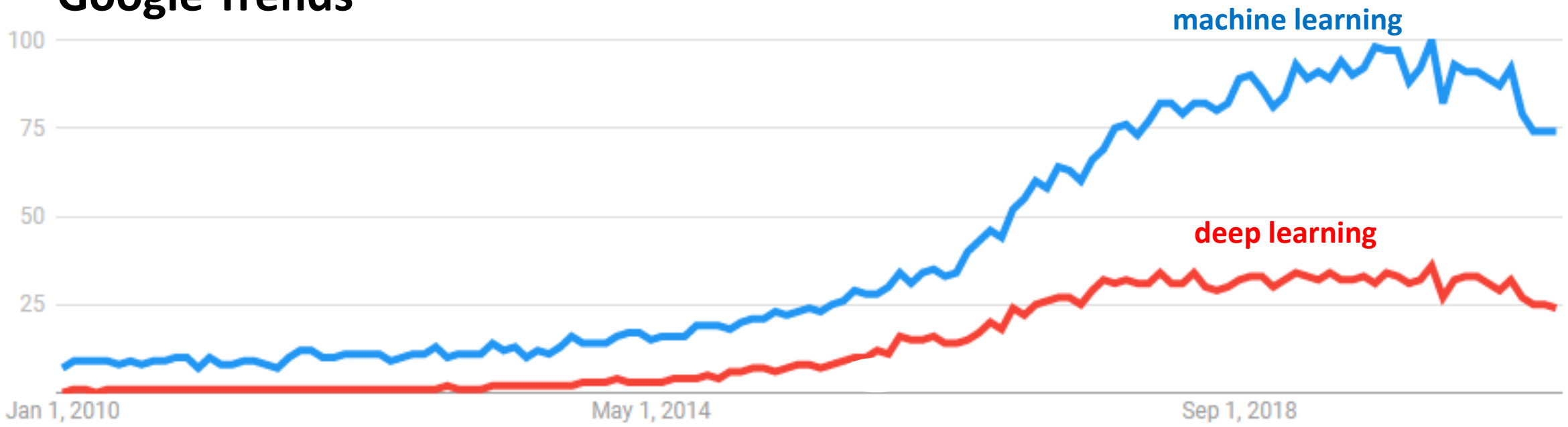
- AI technologie brengt veel voordelen.
- Beloftevolle technologie vergt ook inperken van negatieve aspecten.
- Er zijn tientallen (high level) richtlijnen op ethisch, juridisch en sociaal vlak.



- En nu?
- Wat is eigenlijk het probleem? Verantwoordelijkheid? Voor wat?

Machine Leren: de AI Succesmotor

Google Trends



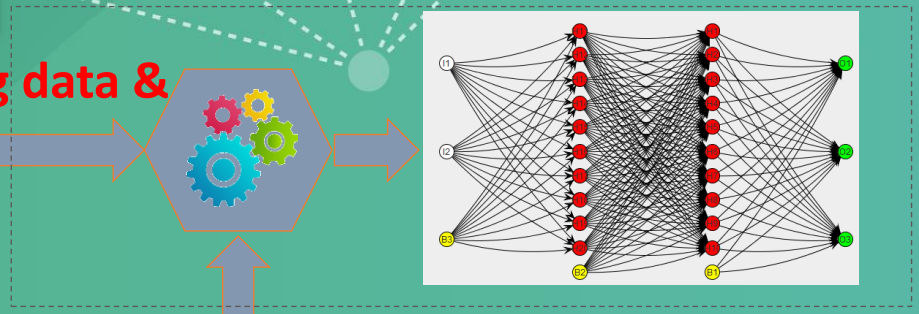
'leervermogen op basis van veel data'

Machine Leren: de AI Succesmotor

‘apple’

LEARNING / TRAINING PHASE

training data & labels



criterion

‘pear’

Model:

1. De relevante wereld kennis

2. Relevante criteria voor prestatie

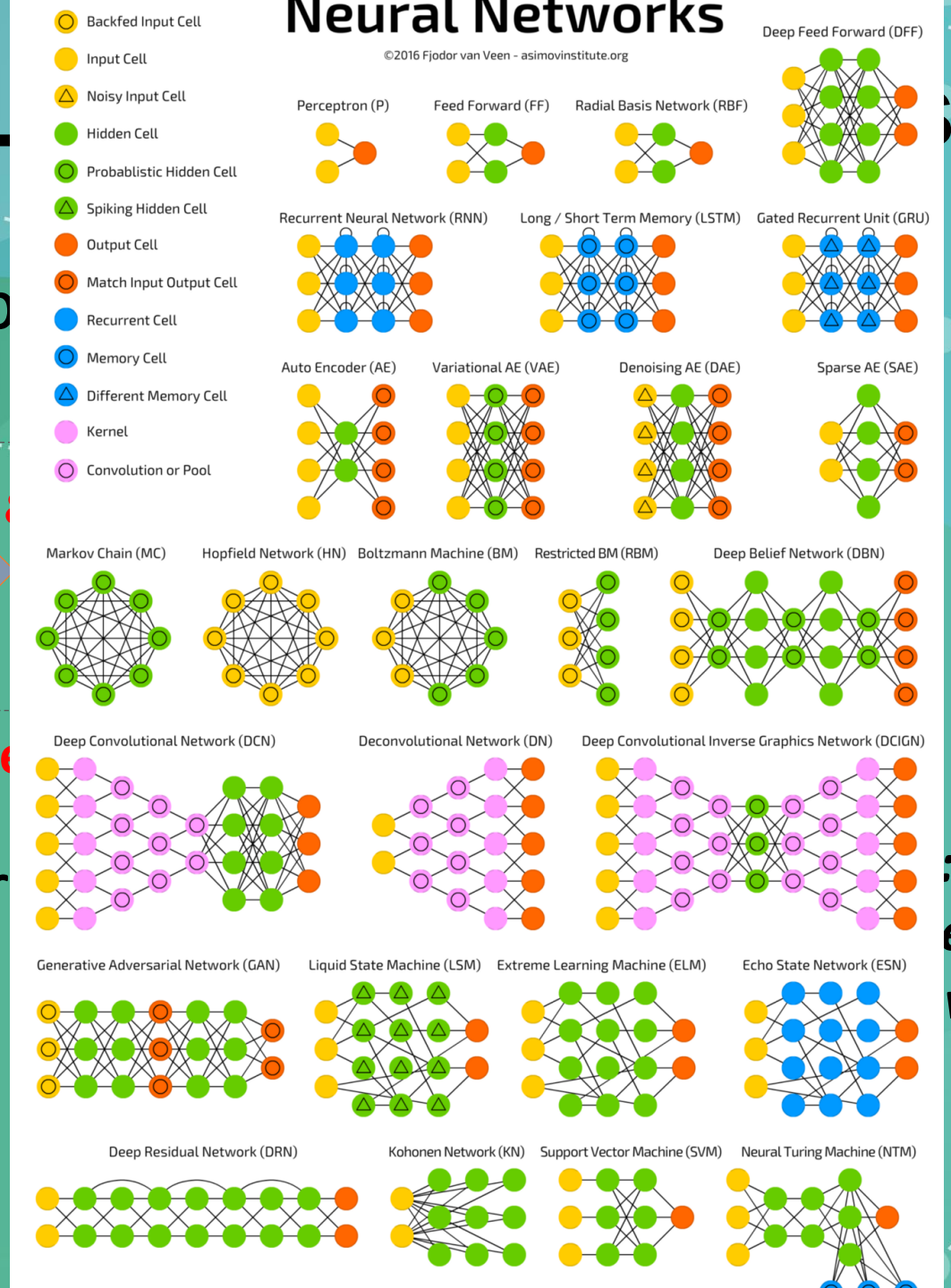
Machine L smotor

'app

training data
labels

crite

'pear



:

elevante wereld kennis

vante criteria voor prestatie

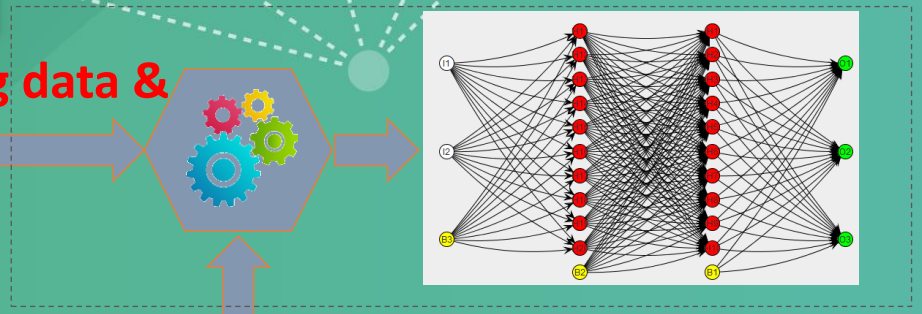
Machine Leren: de AI Succesmotor

‘apple’

LEARNING PHASE

training data & labels

‘pear’



Model:

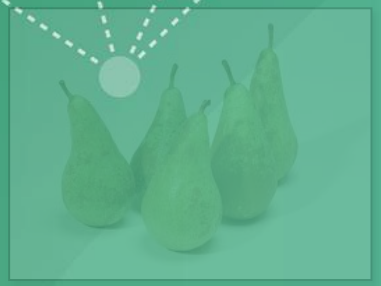
1. De relevante wereld kennis

2. Relevante criteria voor prestatie

Machine Leren: de AI Succesmotor

‘apple’

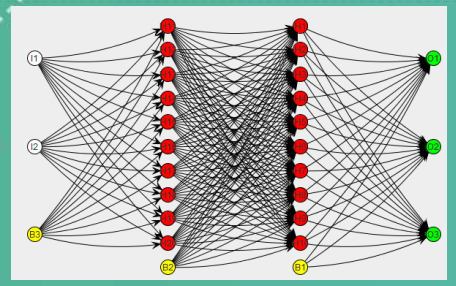
training data &
labels

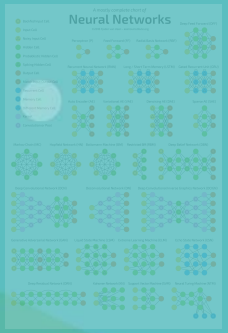


‘pear’

criterion

LEARNING PHASE





APPLICATION PHASE

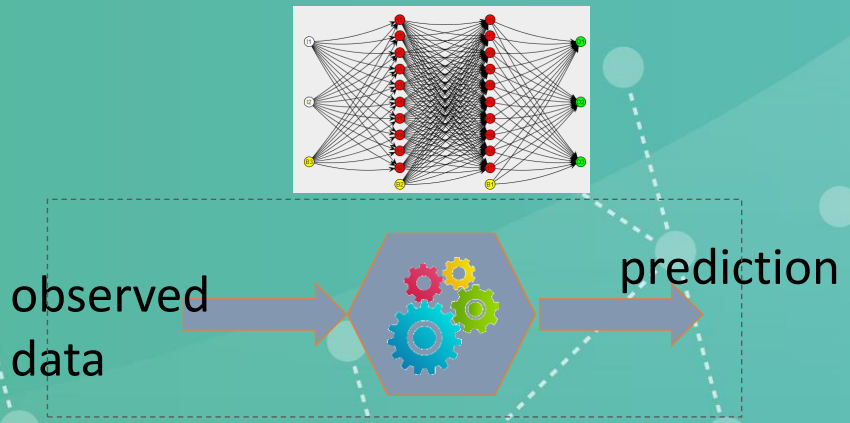
this is a … ?

observed
data

prediction

‘pear’

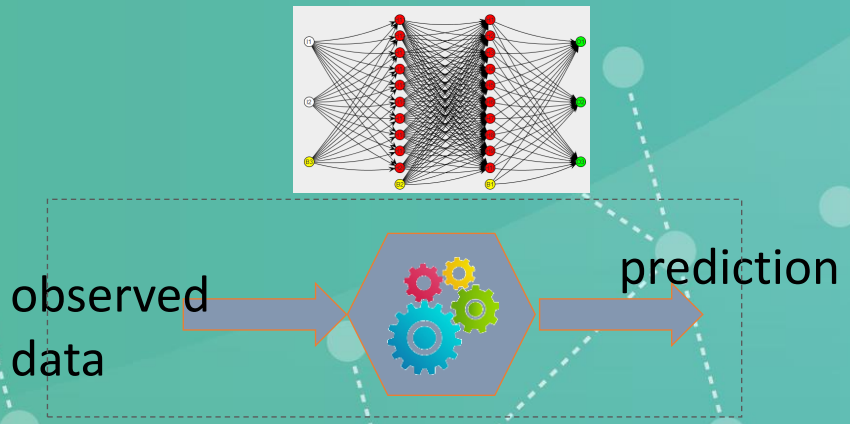
		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		

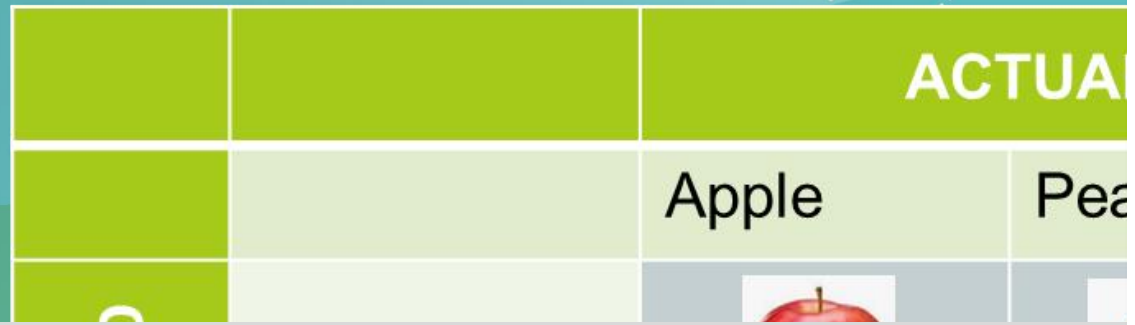


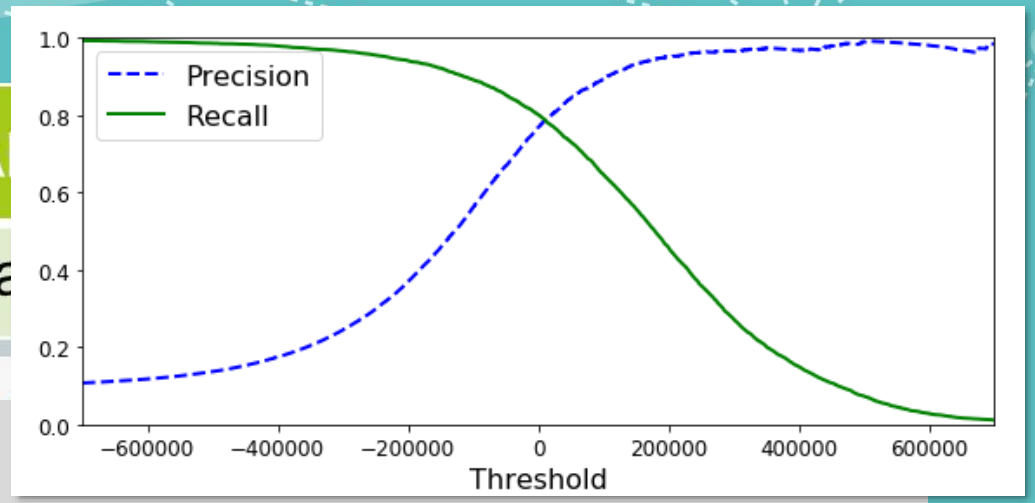
		ACTUAL	
		Apple	Pear
PREDICTED	Apple		
	Pear		

Criterion:

False positives
False negatives
Precision
Recall
F1 measure







Antwoord 1: AI / Algoritmen maken ‘altijd’ fouten

➔ Zijn alle fouten even erg (gebruikt criterium)?

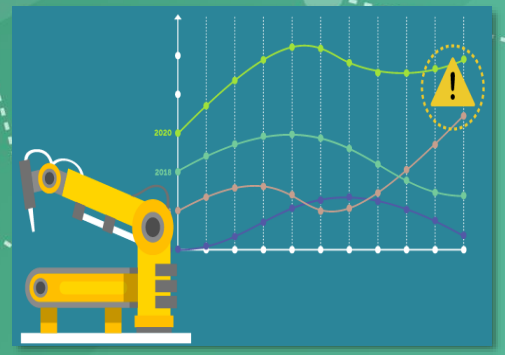
➔ Kunnen wij (of AI) weten dat er een fout gemaakt wordt?

➔ Kunnen we begrijpen waarom (explainable, transparent)?

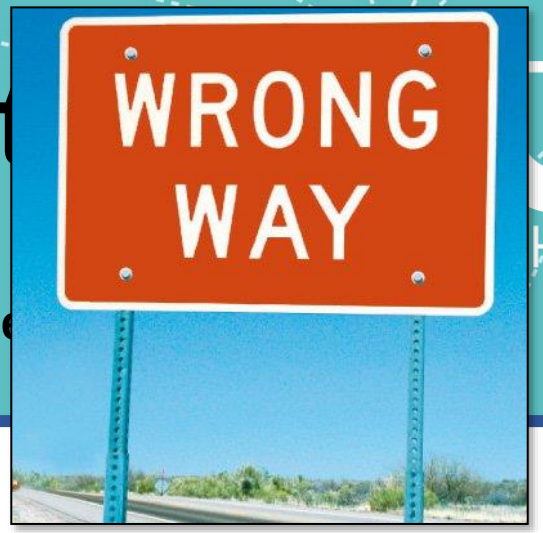
➔ Wie is verantwoordelijk (in de keten van betrokkenen)?

➔ ODD: operational design domain.

Data? Criterium? Errors? Impact?

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

Data? Criterium? Errors? Impact



	Predictive maintenance	Life expectancy prediction	Job suitability prediction	Se
Advantages	Maintenance planning	Informed treatment decisions	High volume selection	Security, marketing
Risks	Reduced value of expertise	Root cause unclarity	Reduced self presentation	Stigmatization

Framing and Formalism trap

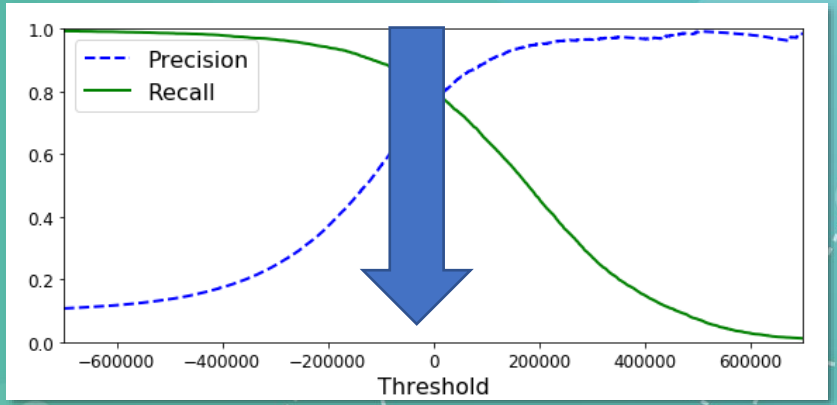
- Formulate goal of the system as an **optimization criterion** (AI-fication).
- What are you **optimizing** for? How much of the world does the data/model capture? (reward hacking and tunnel vision).

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

‘Unlikely …’

	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

UNFAIR!

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

WHAT IS FAIR?

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

Antwoord 2: Vertrouwen op alleen numerieke criteria is problematisch

➔ Slechts dát wordt geoptimaliseerd.

➔ Onverwachte neveneffecten (reward hacking).

➔ Normen, waarden, wetten versus numeriek criterium.

➔ Startpunt is dus: het ideale criterium bestaat niet.

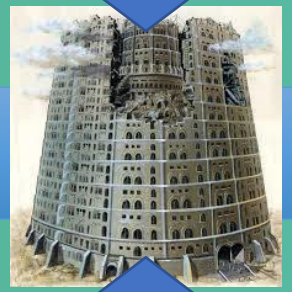
➔ Maar je hebt er wel één nodig. Wie is verantwoordelijk?

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, …

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.

Captures 'criterium issue'

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.

Captures 'criterium issue'

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.

Captures 'errors issue'

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, …

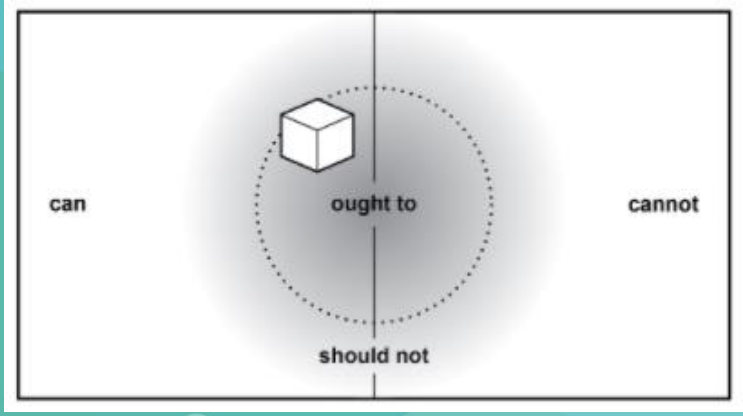
Autonomous System for Good

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, …

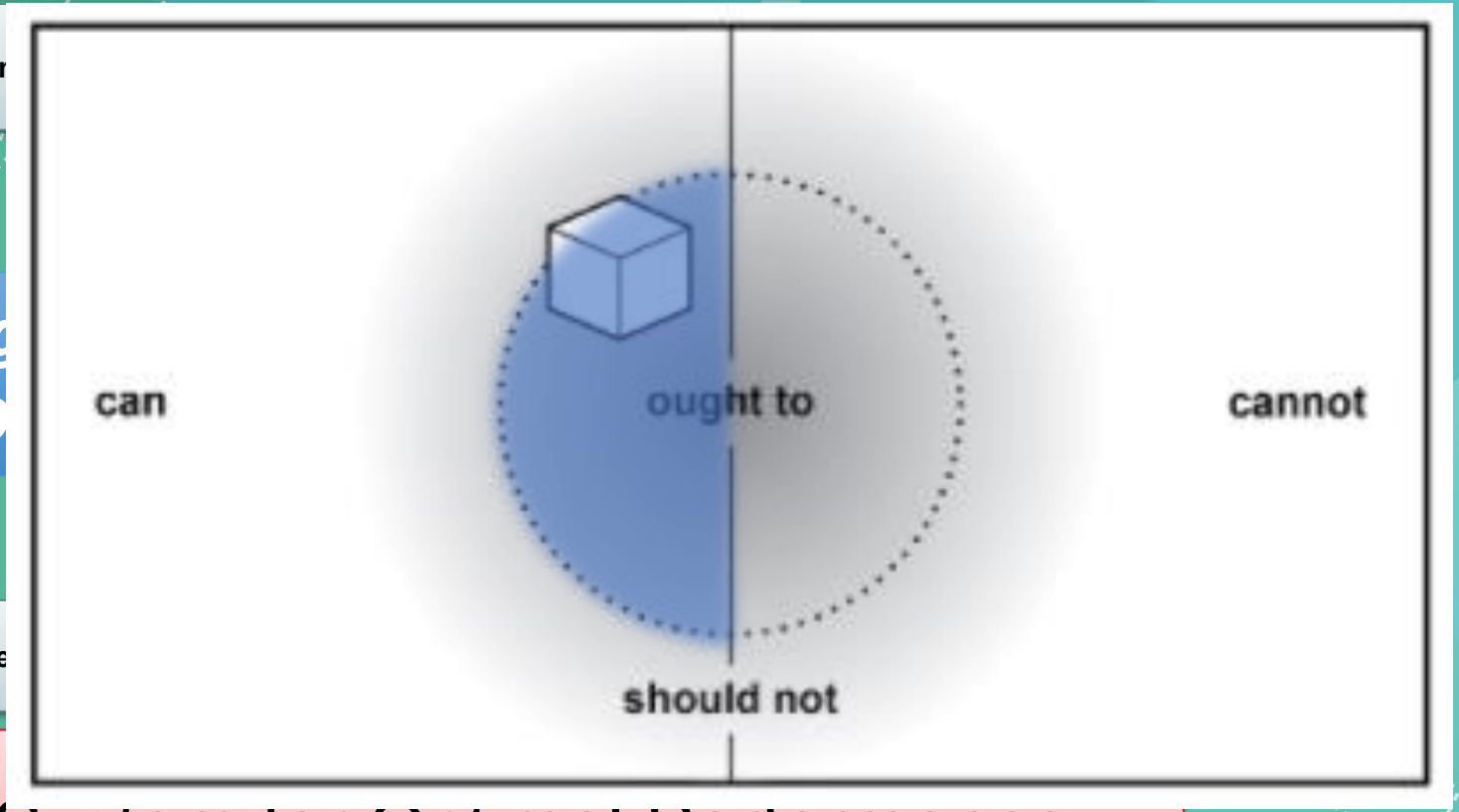
Autonomous System for Good

Law and values: autonomy, beneficence, justice

Instr

Responsibilities

Methodolog
D



Deep le

Autonomous System for Optimal Performance

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

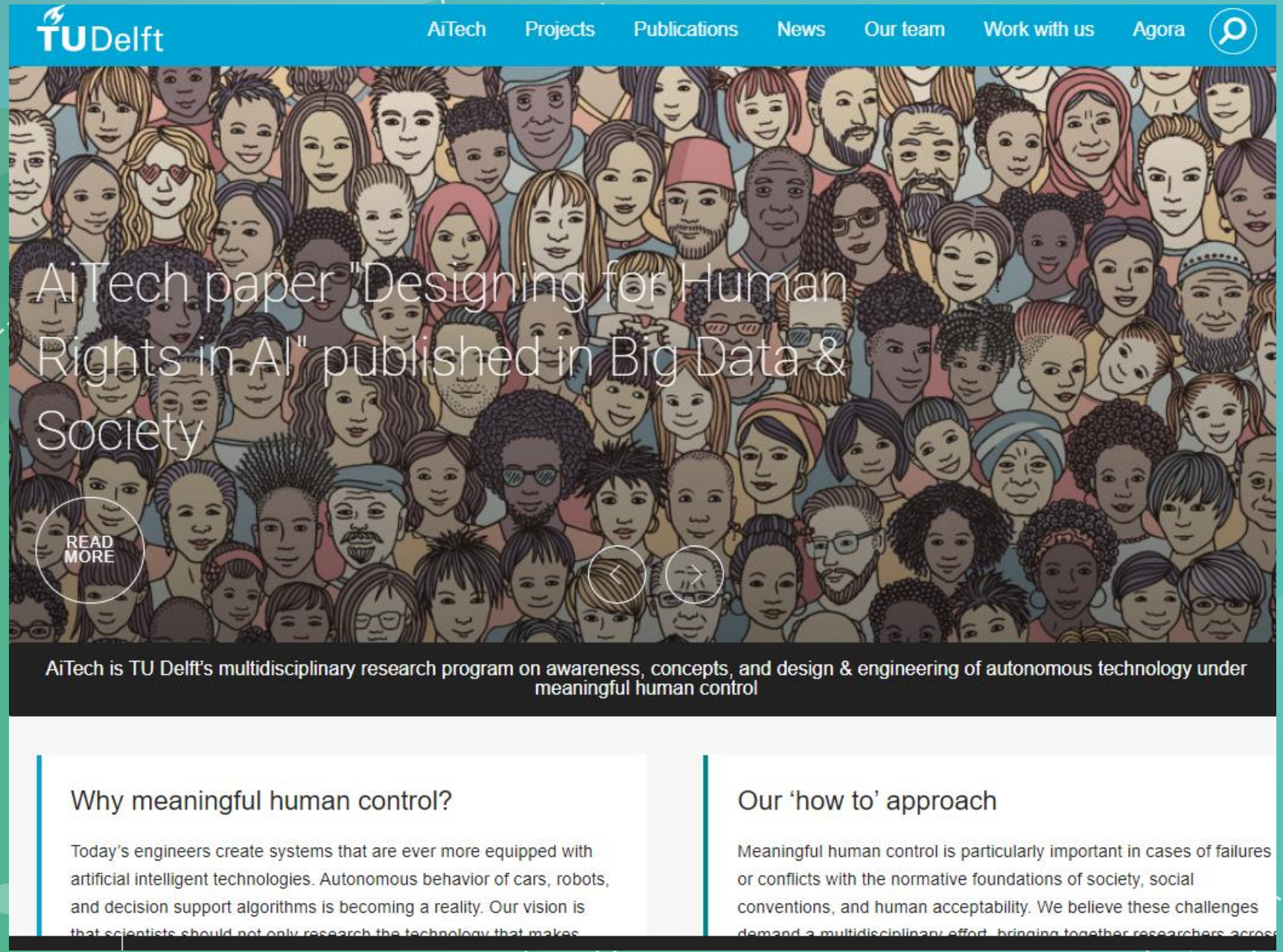
		Qualified person	Not-qualified person
Inside M-ODD	Invited for interview	75	1
Outside M-ODD	Human action	23	29
Inside M-ODD	Not-invited for interview	2	70

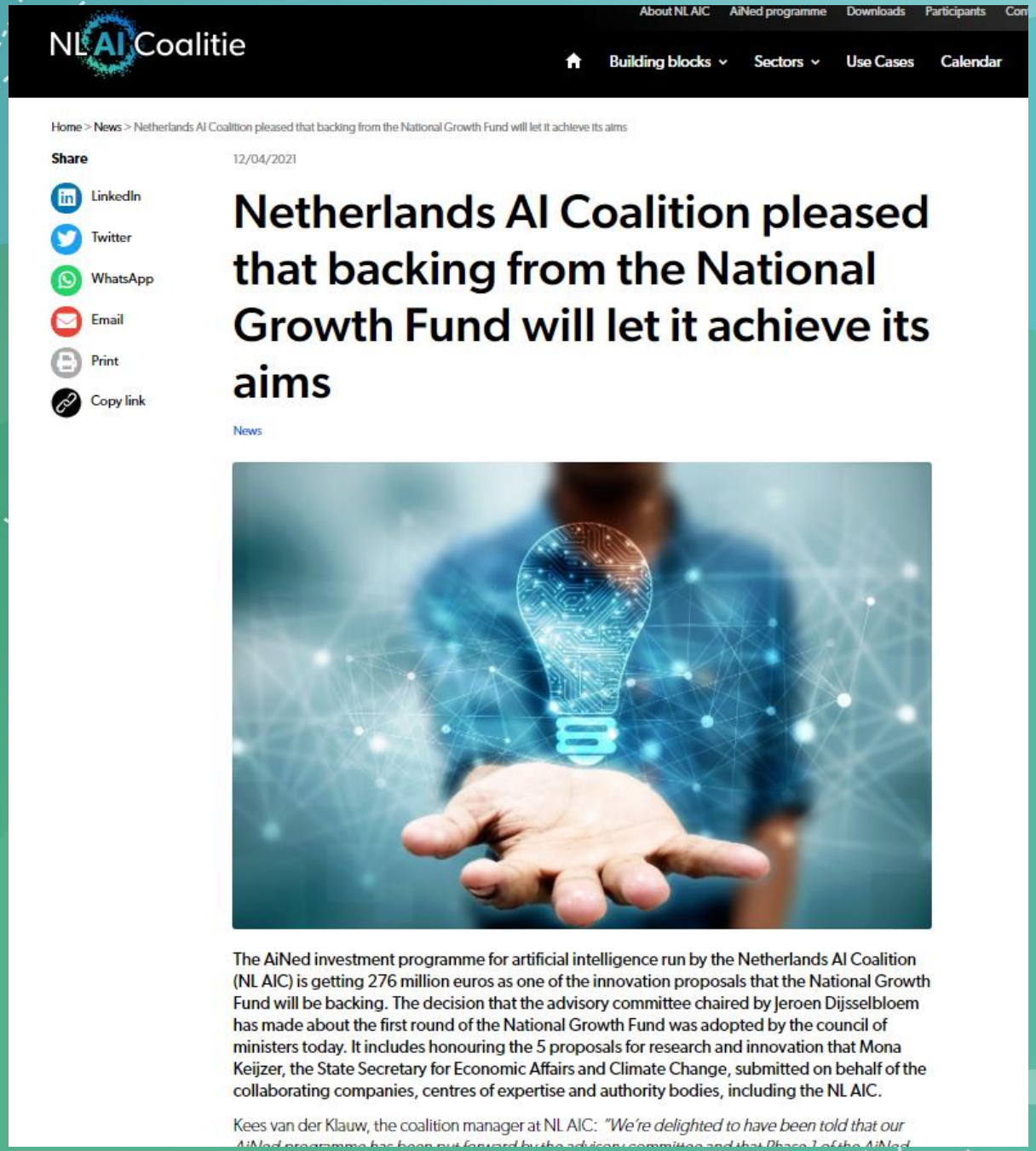
Properties to Realize

- 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.*

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





NL AI Coalitie

About NL AIC | AiNed programme | Downloads | Participants | Con

Building blocks | Sectors | Use Cases | Calendar

Home > News > Netherlands AI Coalition pleased that backing from the National Growth Fund will let it achieve its aims

Share

LinkedIn
Twitter
WhatsApp
Email
Print
Copy link

12/04/2021

Netherlands AI Coalition pleased that backing from the National Growth Fund will let it achieve its aims

News

The AiNed investment programme for artificial intelligence run by the Netherlands AI Coalition (NL AIC) is getting 276 million euros as one of the innovation proposals that the National Growth Fund will be backing. The decision that the advisory committee chaired by Jeroen Dijsselbloem has made about the first round of the National Growth Fund was adopted by the council of ministers today. It includes honouring the 5 proposals for research and innovation that Mona Keijzer, the State Secretary for Economic Affairs and Climate Change, submitted on behalf of the collaborating companies, centres of expertise and authority bodies, including the NL AIC.

Kees van der Klauw, the coalition manager at NL AIC: *"We're delighted to have been told that our*

- **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