AI: dreiging of gift voor security?

Dr Stefan Leijnen
Lector AI

Vs

Dr Sieuwert van Otterloo CISA CIPP/E
IT auditor and AI onderzoeker

Agenda van vanavond

1. AI als gift: Wat kan AI?
2. AI als dreiging: problemen door AI
3. Debat: Hoe willen we omgaan met AI?

Sprekers

Dr. Sieuwert van Otterloo is IT-expert en directeur van adviesbureau ICT Institute. Hij geeft advies over IT-strategie, informatiebeveiliging en privacy en voert regelmatig audits uit. Daarnaast geeft hij les aan de Vrije Universiteit en is AI onderzoeker bij de Hogeschool Utrecht.
Zie www.softwarezaken.nl

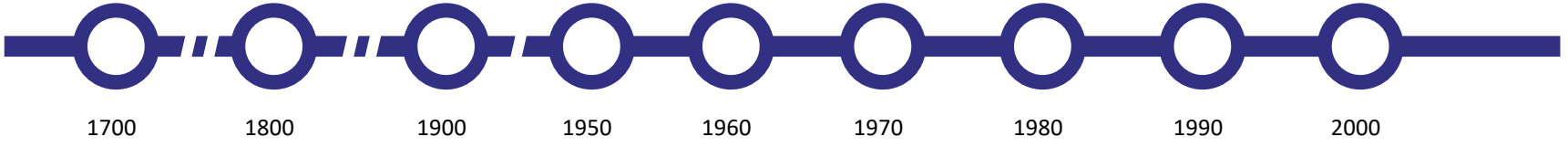
Dr. Stefan Leijnen is lector van het lectoraat Artificial Intelligence bij de Hogeschool Utrecht, waar hij onderzoek doet naar toepassingen van artificial intelligence en data-gedreven innovatie. Mensgerichte waarden als transparantie, vertrouwen, creativiteit en autonomie staan daarbij centraal.

Zie https://www.hu.nl/onderzoek/artificial-intelligence

Agenda van vanavond

1. AI als gift: Wat kan AI?
2. AI als dreiging: problemen door AI
3. Debat: Hoe willen we omgaan met AI?

AI is al heel oud ….



AI als science fiction

AI experimenten

Uitvinding computer
en theorieën

Klassieke AI

Moderne AI /
Datagedreven AI

… maar is deze eeuw actueel geworden

Veel nieuwe mogelijkheden door verbeteringen in technologie en algoritmes

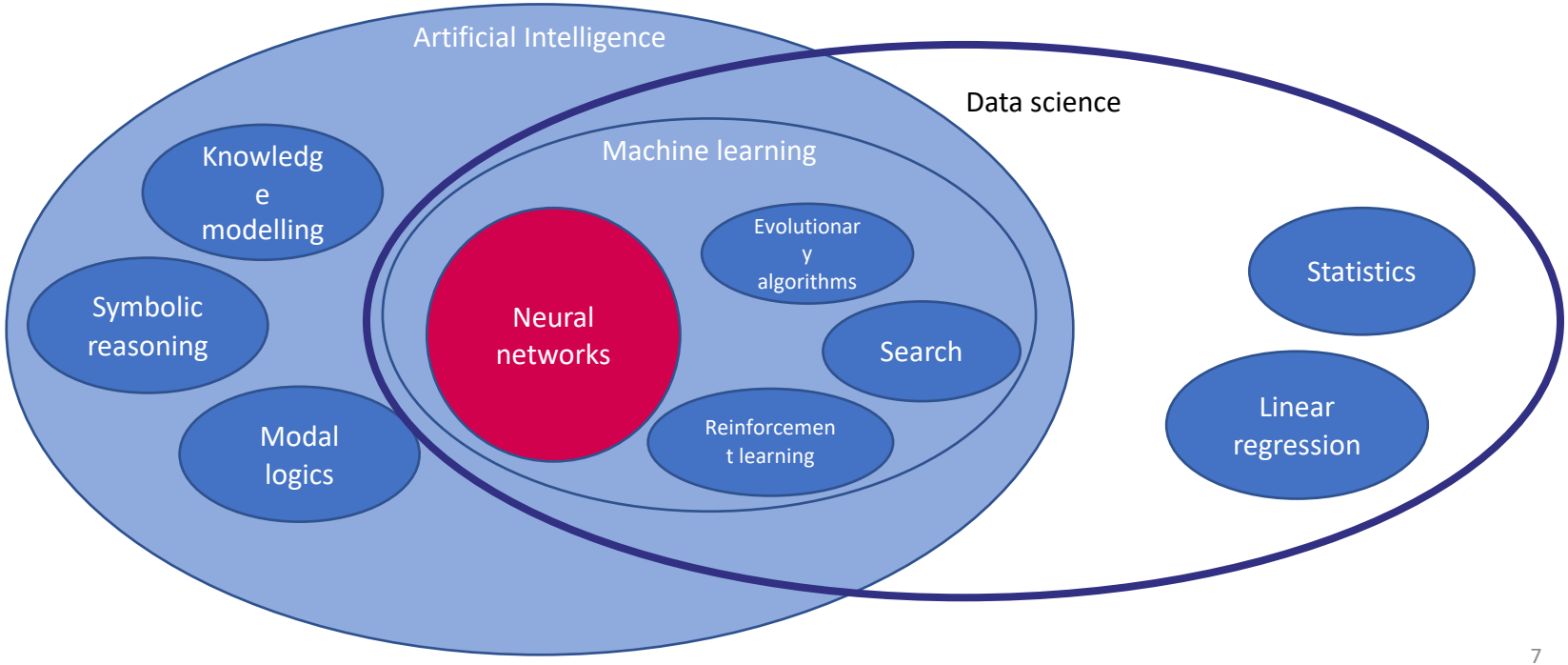
amazon

TESLA

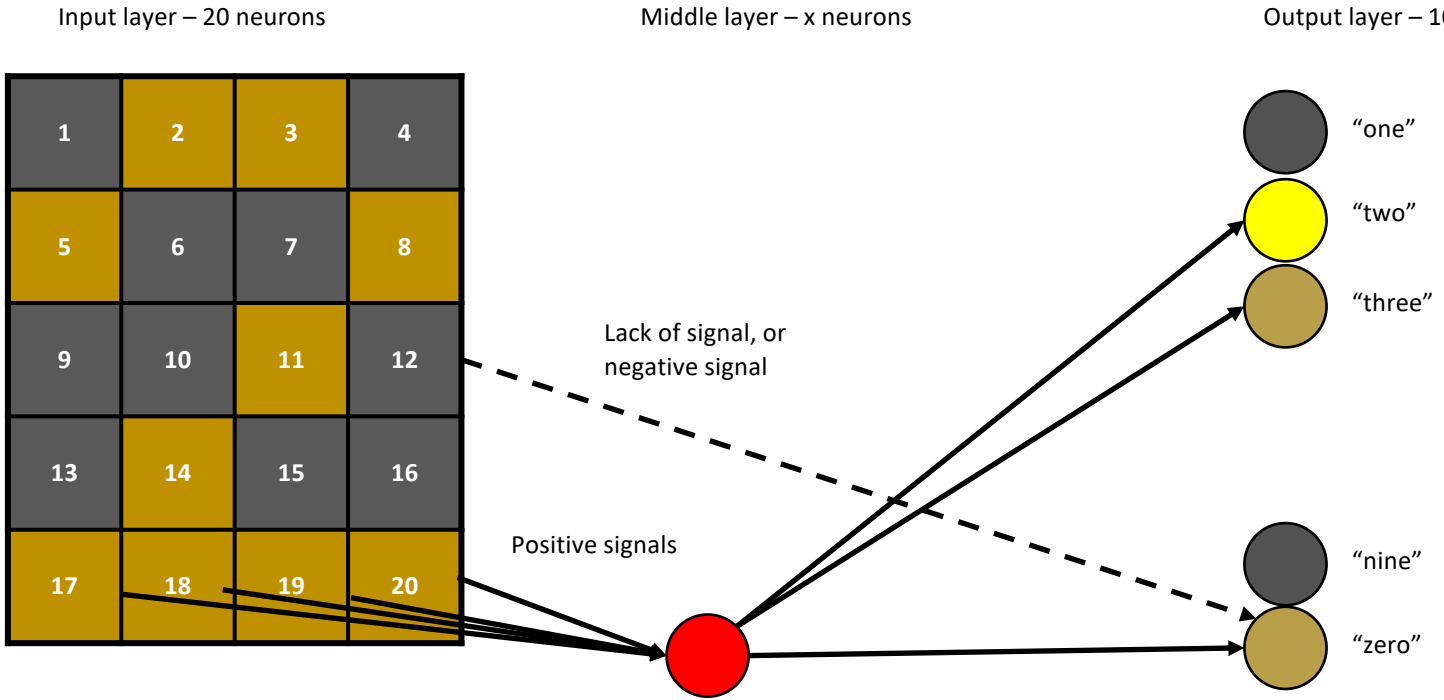
Veel bedrijven die extra veel data verzamelen omdat data veel waard aan het worden is (Tesla, Amazon, Netflix, Microsoft)

Veel risico's en problemen (filosofisch, technisch, juridisch, praktisch)

2022: AI is een groot en divers onderzoeksgebied



Voorbeeld: neuraal netwerk voor beeldherkenning



2000 … 2022: datagedreven AI

Klassieke IT

Kenmerken

- De kennis wordt door de developers vastgelegd
- Gedrag van systemen is voorspelbaar

Klassieke AI

- Niet-programmeurs kunnen kennisregels modelleren
- De kennis van het systeem is eenvoudiger aanpasbaar

Datagedreven AI

- Heeft heel veel data nodig (heel veel)
- Ontdekt zelf patronen die experts misschien niet kennen
- Kan leiden tot onverwachte fouten: niemand weet welke regels worden gebruikt

Trainen van AI

- Mensen of computers verzamelen foto’s
- Experts voegen tags toe van concepten die geleerd moeten worden
- De computer leert op basis van de tags om zelf foto’s te categoriseren

tuin

trap

modern

Geen bezoekers

Geen schade

Vluchtweg vrij

Kans 1 : AI voor intrustion detection

Klassieke firewalls netwerktools zijn regelgebaseerd:

- Verkeer dat niet aan de regels voldoet (CFO op vakantie) wordt geweigerd
- Verkeer dat aan de regels voldoet wordt toegelaten (hackers op locatie))

AI intrusion detection leert normal verkeer te voorspellen en geeft waarschuwingen bij afwijkend gedrag:

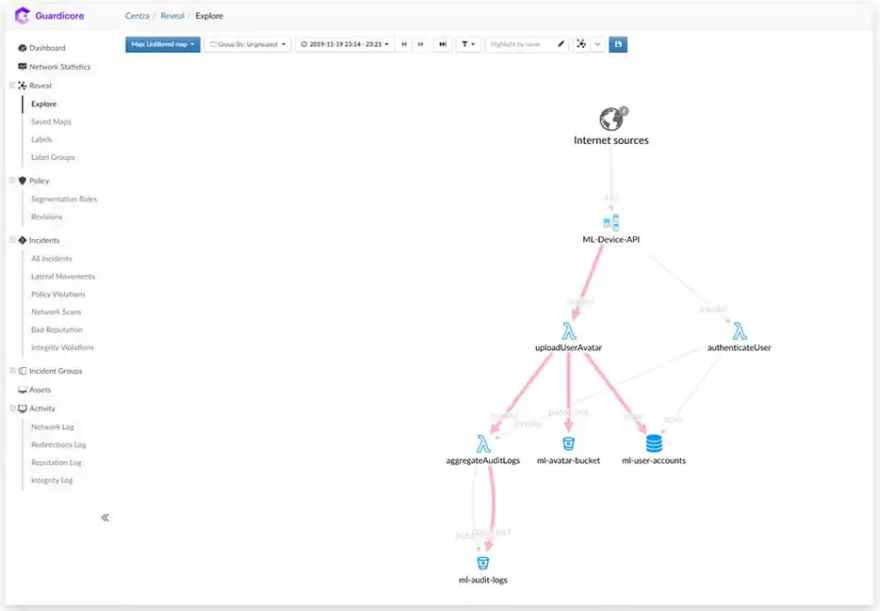
- Minder waarschuwingen
- Bij afkeuren van melding kan er gelijk worden gehertraind

Voorbeeld: Guardicore

Centra's AI-powered segmentation reduces the time it takes to create a segmentation policy for a new or existing application by making it easier to label assets and create the matching rules for them.

…

Our AI-based algorithm is capable of 'learning' tens of thousands of applications and millions of flows, allowing us to provide: 1) tailored policy templates based on the customer's assets and 2) automatic labels tailored to the customer's environment.



https://www.akamai.com/blog/security/ai-powered-segmentation-cloud-native-security

Andere voorbeelden

- Nieuwe authenticatie: gezichtsherkenning en biometrics
- Herkennen van fraude-gevoelige transacties
- Filteren van gevoelige content, niet-passende foto’s

Conclusies

- AI is heel geschikt voor het weg-automatiseren van beslisproblemen: De computer kan dit, met de juiste data, veel sneller en beter dan mensen
- Door AI verdwijnt het monotone security-werk en kan de CISO zich richten op de trends en strategische ontwikkelingen
- AI vervangt geen security-experts maar maakt hen veel effectiever

Wat voor AI verwacht je binnen vijf jaar te gebruiken in je organisatie

- Chatbots
- Adaptieve trainingen
- Zelfconfigurerende systemen
- Hack-detectie
- Beeldherkenning
- Video-analyse

https://www.menti.com/al4wp25h9ajf

Agenda van vanavond

1 AI als gift: Wat kan AI?

2 AI als dreiging: problemen door AI

3 Debat: Hoe willen we omgaan met AI?

Dreiging 1: Deepfakes

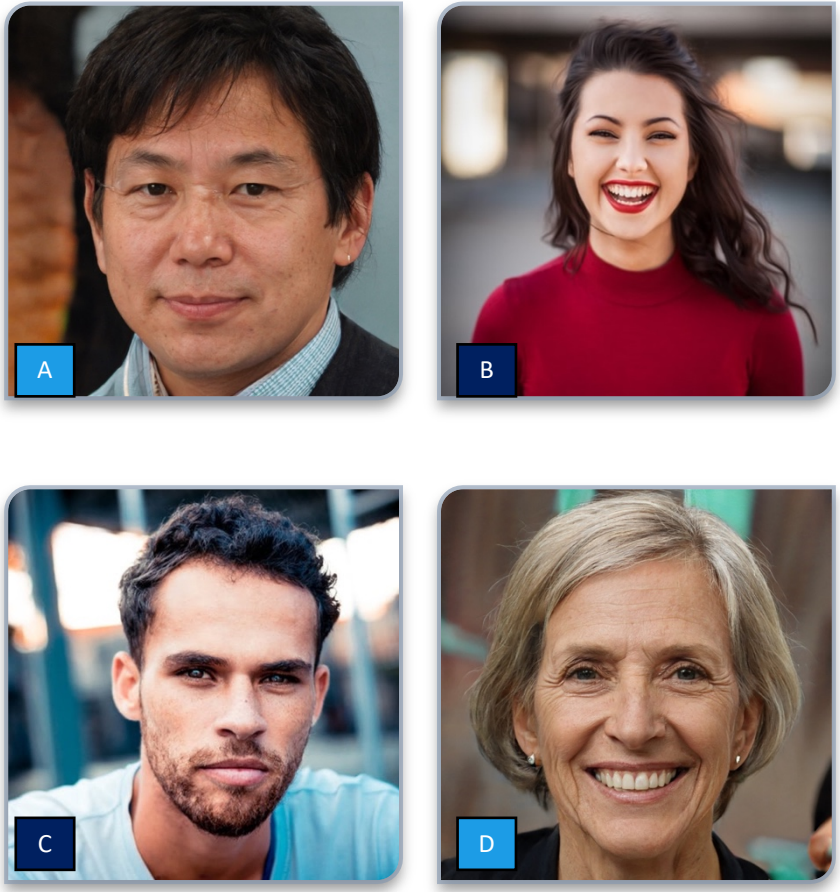
Computers kunnen realistische beelden maken, zogenaamde ’deepfakes’. Het wordt steeds moeilijker deze beelden van echt te onderscheiden.
De foto hiernaast is gegenereerd met een generative adversarial neural network: Het ene netwerk leert een bestaand AI systeem voor de gek te houden.

Deepfake test

Wat denken jullie?

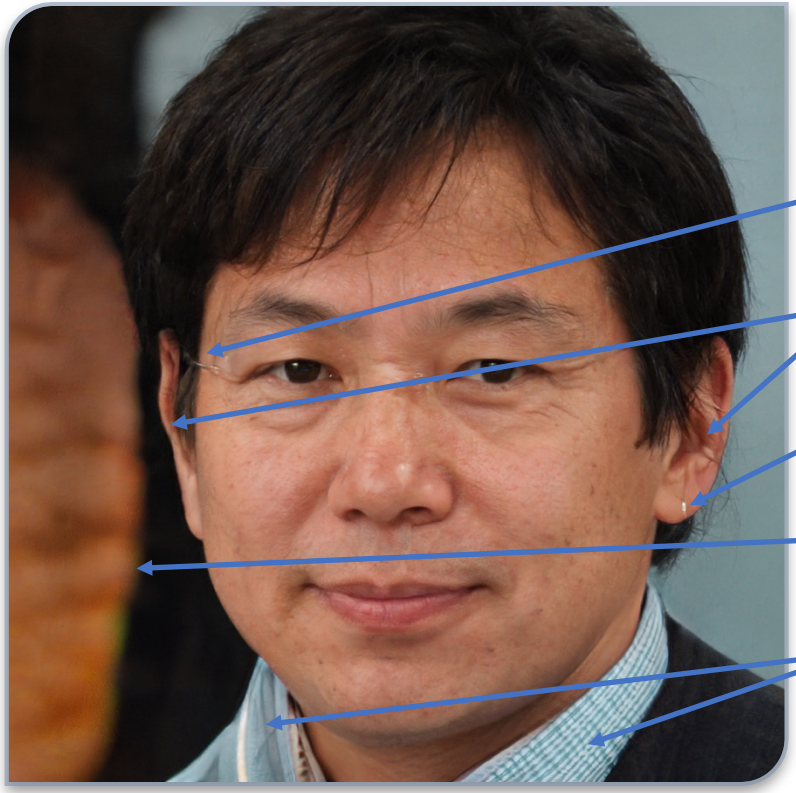
A en D zijn echt
B,C zijn deepfakes

B en C zijn echt
A,D zijn deepfakes



https://thispersondoesnotexist.com
Onderzoek door Sebastiaan Spanjer

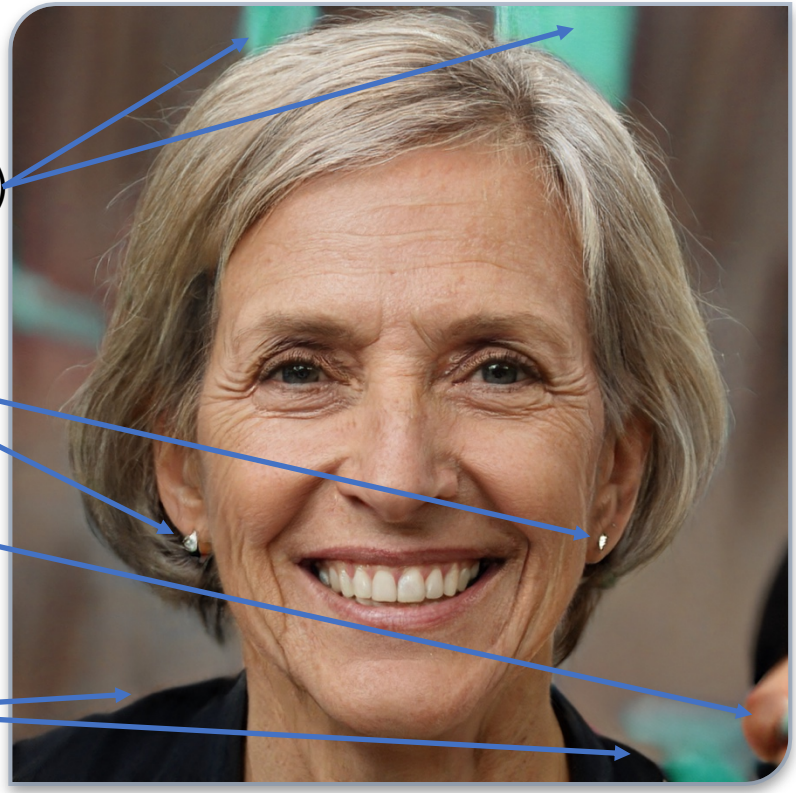
How to recognize a generated photo? (1/2)



- Residue of glasses (general odd occurrence)
- Unevenly sized ears (symmetry)
- Asymmetrical earring(s) (symmetry)
- Weird shape in background (background)
- Asymmetrical/ strangely colored clothes (symmetry)

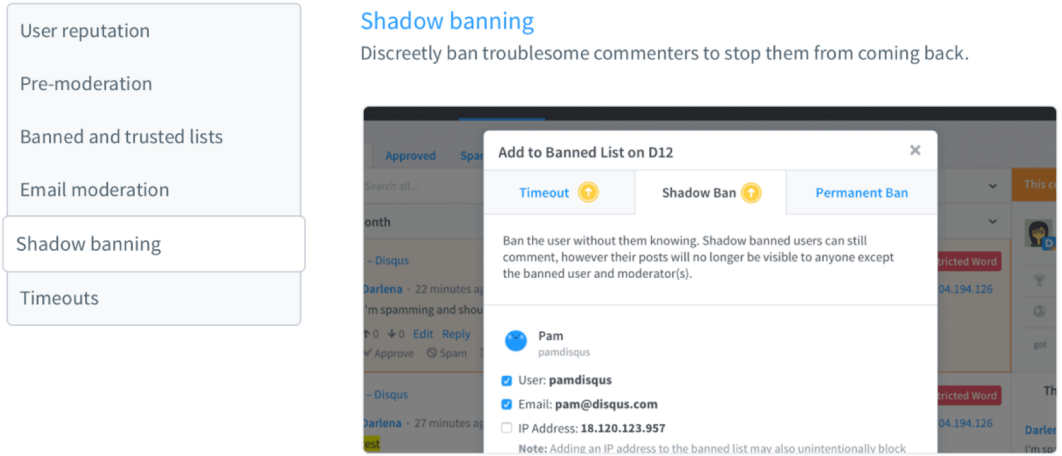
How to recognize a generated photo? (2/2)

FAKE



- Weird shapes in background (background)
- Asymmetrical earrings (symmetry)
- Weird blurred artefact in background (general odd occurrence)
- Uneven shoulder height (symmetry)

Dreiging 2: gebrek aan transparantie



Websites gebruiken AI om risico's in te schatten, en uiteindelijk besluiten te nemen, bijvoorbeeld bannen gebruikers. De modellen hiervoor zijn complex. Platforms zelf weten niet altijd hoe het werkt.

Het gebruik is niet altijd transparant. Bij 'shadowbanning' weet een gebruiker niet dat hun input genegeerd wordt. Dit wordt gedaan om klachten te voorkomen.

https://disqus.com/features/engage/

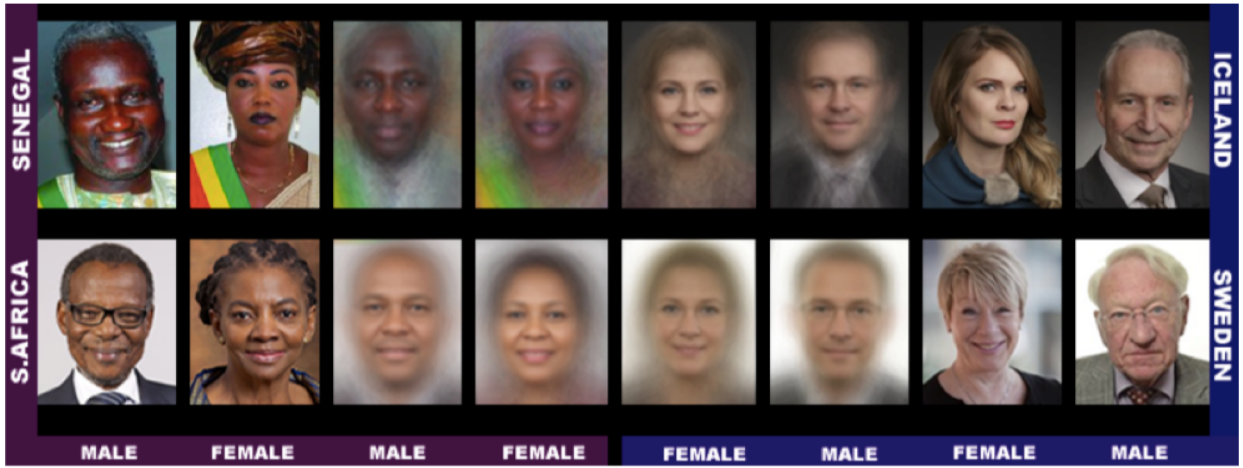
Dreiging 3: bias

Veel AI-systemen hebben ‘bias’.
Bijvoorbeeld:

- De diagnoses voor jongeren en mannen zijn beter dan voor ouderen en vrouwen
- AI-systemen herkennen gezichten van donkere mensen minder goed

Hoeveel bias is toegestaan?

Voorbeeld bias in gezichtsherkenning



Computer programs recognise white men better than black women

Biased training is probably to blame - Feb 17 2018 Economist

..

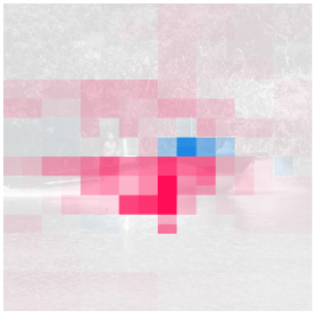
The algorithms involved have, however, long been suspected of bias. Specifically, they are alleged to be better at processing white faces than those of other people.

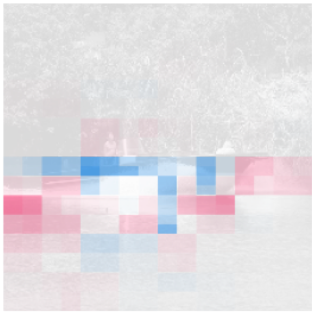
http://gendershades.org

Recht op uitleg

SHAP is een algoritme dat laat zien welke input een algoritme gebruikt om een conclusie te trekken.
In dit voorbeeld zie je welk deel van de foto's is gebruikt.
Er wordt veel onderzoek gedaan naar betere uitlegmethodes, zowel voor onderzoekers als voor de eindgebruikers

De meeste algoritmes zijn helaas niet uitlegbaar





https://shap.readthedocs.io/en/stable/example_notebooks/image_examples/image_classification/Explain%20ResNet50%20using%20the%20Partition%20explainer.html

Conclusies

AI maakt de maatschappij niet veiliger, omdat het nieuwe problemen introduceert:

- Gebrek aan transparantie over AI-besluiten
- Verkeerde besluiten gebaseerd op bias die lasting zijn te herstellen
- AI-gebaseerde aanvallen met behulp van DeepFakes.

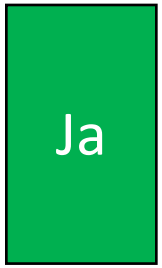
Agenda van vanavond

1 AI als gift: Wat kan AI?

2 AI als dreiging: problemen door AI

3 Debat: Hoe willen we omgaan met AI?

Moeten we AI inzetten voor monitoring?



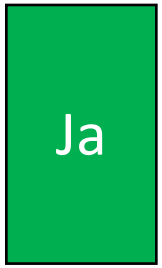
Ja, want

- Steeds meer te monitoren assets
- Complexere aanvallen
- Gebrek aan opgeleide mensen

Nee, want

- AI is niet feilloos
- Fouten moeilijk op te sporen en te herstellen

Leidt AI tot minder werk voor CISO’s?



Ja, want

- Veel CISO-werk is automatiseerbaar

Nee, want

- AI moet ook getraind worden

Moet ‘kwade’ AI worden verboden

Ja

Ja, want

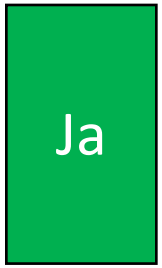
- Deep fakes zijn gevaarlijk
- AI-gebaseerde aanvallen zijn moeilijk te stoppen

Nee

Nee, want

- Technologie is niet goed of kwaad
- Technologie verbieden houdt criminelen niet tegen

Een beslissing, AI of niet, moet altijd uitlegbaar zijn



Ja, want..

… wilt controleren

… moet bezwaar kunnen maken

… je moet opnieuw kunnen proberen

Nee, want ….

… menselijke besluiten zijn ook moeilijk uitlegbaar

… het gaat om de kwaliteit, niet om de uitleg

Moet Nederland meer investeren in AI?

Ja

Ja want AI is de toekomst

Nee

Nee,

…er wordt al genoeg geïnvesteerd

…alleen in verantwoorde AI en betrouwbaarder maken van AI