



# ChatGPT: hoe werkt het?

Een kijkje onder de digitale  
motorkap



# Giulia Bössenecker

- Artificial intelligence aan de Vrije Universiteit Amsterdam
- Data en AI adviseur bij het Rijks ICT Gilde
- Verantwoorde ontwikkeling en inzet van AI bij de Rijksoverheid
- Overheidsbrede visie op generatieve AI, algoritmekader en IAMA pilot



**ChatGPT**



Schrijf een gedicht van vier zinnen over het belang van begrijpen van AI

**INPUT**



In een wereld vol mysterie en schijn,  
Opent begrip voor AI nieuwe deuren wijd en klein.  
Het onthult de kracht van technologie,  
En leidt ons naar een toekomst vol harmonie.

**OUTPUT**



```
394 class OPTPreTrainedModel(PreTrainedModel):
395
396     config_class = OPTConfig
397     base_model_prefix = "opt"
398     supports_gradient_checkpointing = True
399     _no_split_modules = ["OPTDecoderLayer"]
400     _keys_to_ignore_on_load_unexpected = [r"decoder_self_attn.bias"]
401
402     def _init_weights(self, module):
403         std = self.config.init_std
404         if isinstance(module, nn.Linear):
405             module.weight.data.normal_(mean=0.0, std=std)
406             if module.bias is not None:
407                 module.bias.data.zero_()
408         elif isinstance(module, nn.Embedding):
409             module.weight.data.normal_(mean=0.0, std=std)
410             if module.padding_idx is not None:
411                 module.weight.data[module.padding_idx].zero_()
```



# GPT = Generative Pre-trained Transformer

---

## Attention Is All You Need

---

**Ashish Vaswani\***  
Google Brain  
avaswani@google.com

**Noam Shazeer\***  
Google Brain  
noam@google.com

**Niki Parmar\***  
Google Research  
nikip@google.com

**Jakob Uszkoreit\***  
Google Research  
usz@google.com

**Llion Jones\***  
Google Research  
llion@google.com

**Aidan N. Gomez\* †**  
University of Toronto  
aidan@cs.toronto.edu

**Lukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention

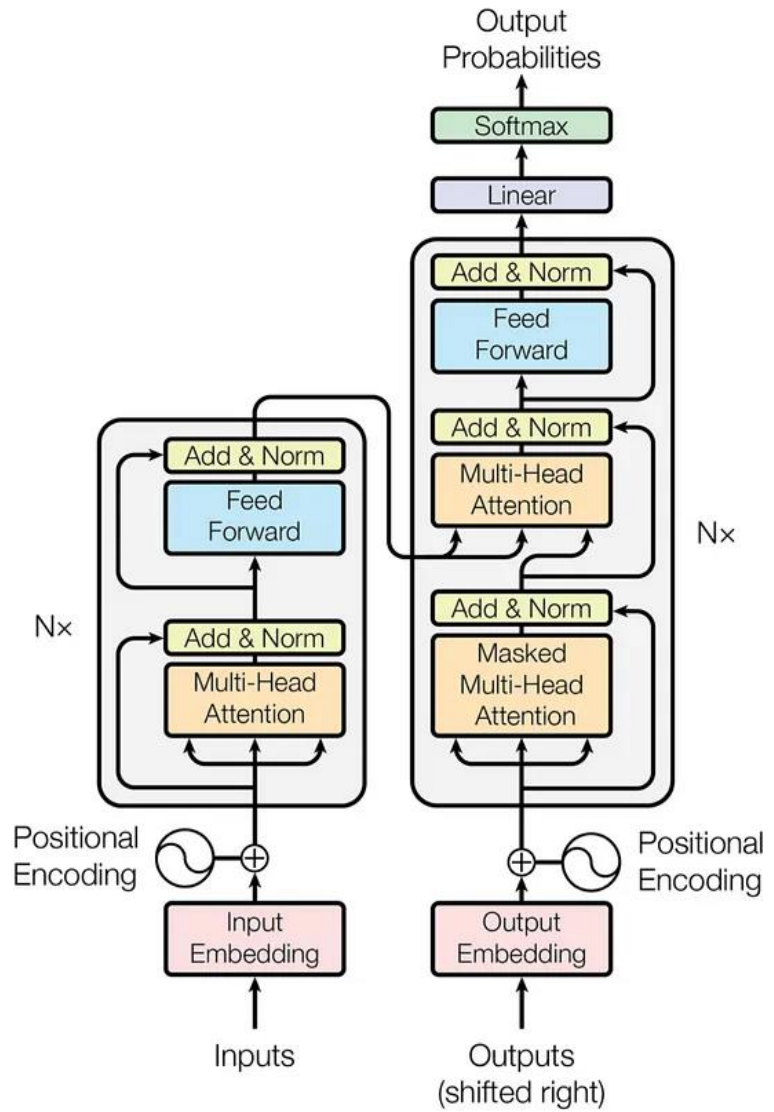


Figure 1: The Transformer - model architecture.

- Rekenstappen en parameters
  - Rekenstappen bedacht door ontwikkelaars
  - Parameters ontwikkeld door training
  - Bouwblok voor Large Language Models
- 
- GPT-4 heft 175 miljard parameters

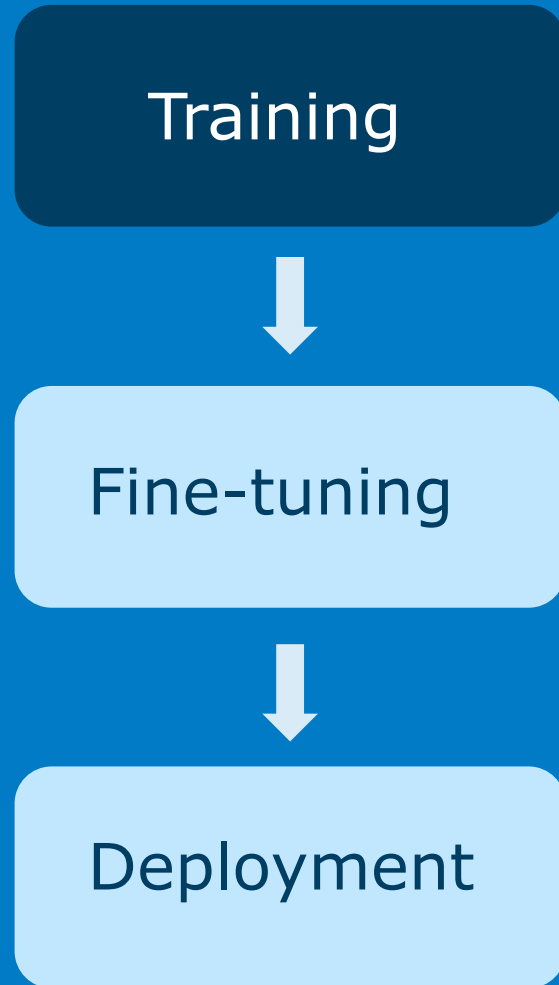


# Self-attention

Ik open een rekening bij de bank

Vanavond ga ik lekker op de bank zitten





- ~ 50 terabytes aan data voor GPT 3.5
- Patronen herkennen in de data
- Parameters verschuiven
- Miljarden herhalingen
- Weinig menselijke tussenkomst
- GPT-4



Training



Fine-tuning



Deployment

- Toespitsen op een doel
- Eventuele vangrails
- Reinforcement Learning from Human Feedback (RLHF)
- ChatGPT



Training



Fine-tuning



Deployment

- Klaar voor gebruik
- Statistisch model
- ChatGPT maar ook Midjourney, DALL-E en meer



**G** Schrijf een gedicht van vier zinnen over het belang van begrijpen van AI



In een wereld vol mysterie en schijn,  
Opent begrip voor AI nieuwe deuren wijd en klein.  
Het onthult de kracht van technologie,  
En leidt ons naar een toekomst vol harmonie.

**G** Schrijf een kort gedicht van vier zinnen over het belang van begrijpen van AI



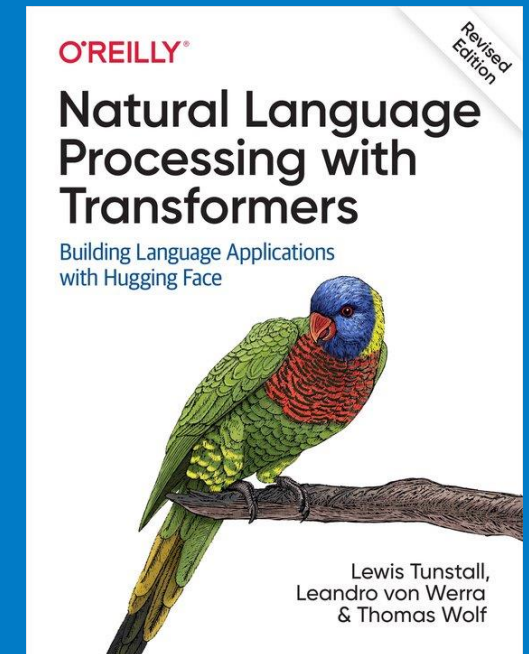
In een wereld vol wonderen, vol AI-kracht,  
Is begrijpen van deze kunst een ware schat.  
Met kennis en inzicht, de sleutel in hand,  
Leidt begrip ons naar een toekomst zo grand.



# Dank voor jullie aandacht!

## Bronnen

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. " O'Reilly Media, Inc."
- Kublik, S., & Saboo, S. (2022). *GPT-3*. O'Reilly Media, Incorporated.



# Randvoorwaarden voor technologie



**Iris Muis**  
Teamleider Data School



## 8 jaar ervaring met digitale ethiek bij overheden

- Begeleiden van gesprekken over de (on)wenselijkheid van technologie;
- Cursussen en trainingen voor ambtenaren en raadsleden;
- Onderzoek naar randvoorwaarden voor het verantwoord toepassen van technologie;
- Ontwikkeling van instrumenten die hierbij helpen.

Deugt het en doet het deugd?



Kan het, mag het en willen we het?

# Kan het, mag het en willen we het?

Technisch

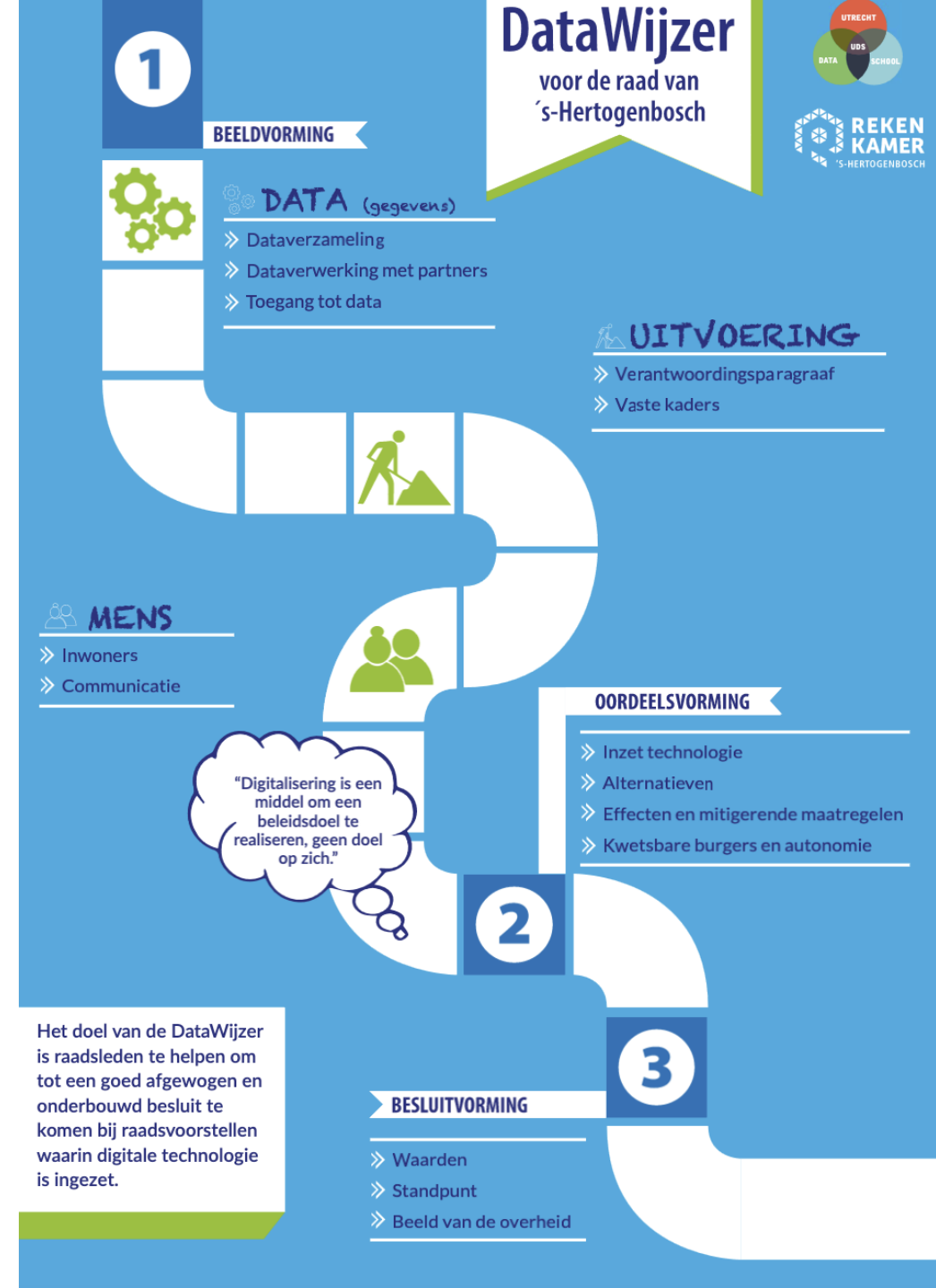
Juridisch

Ethisch

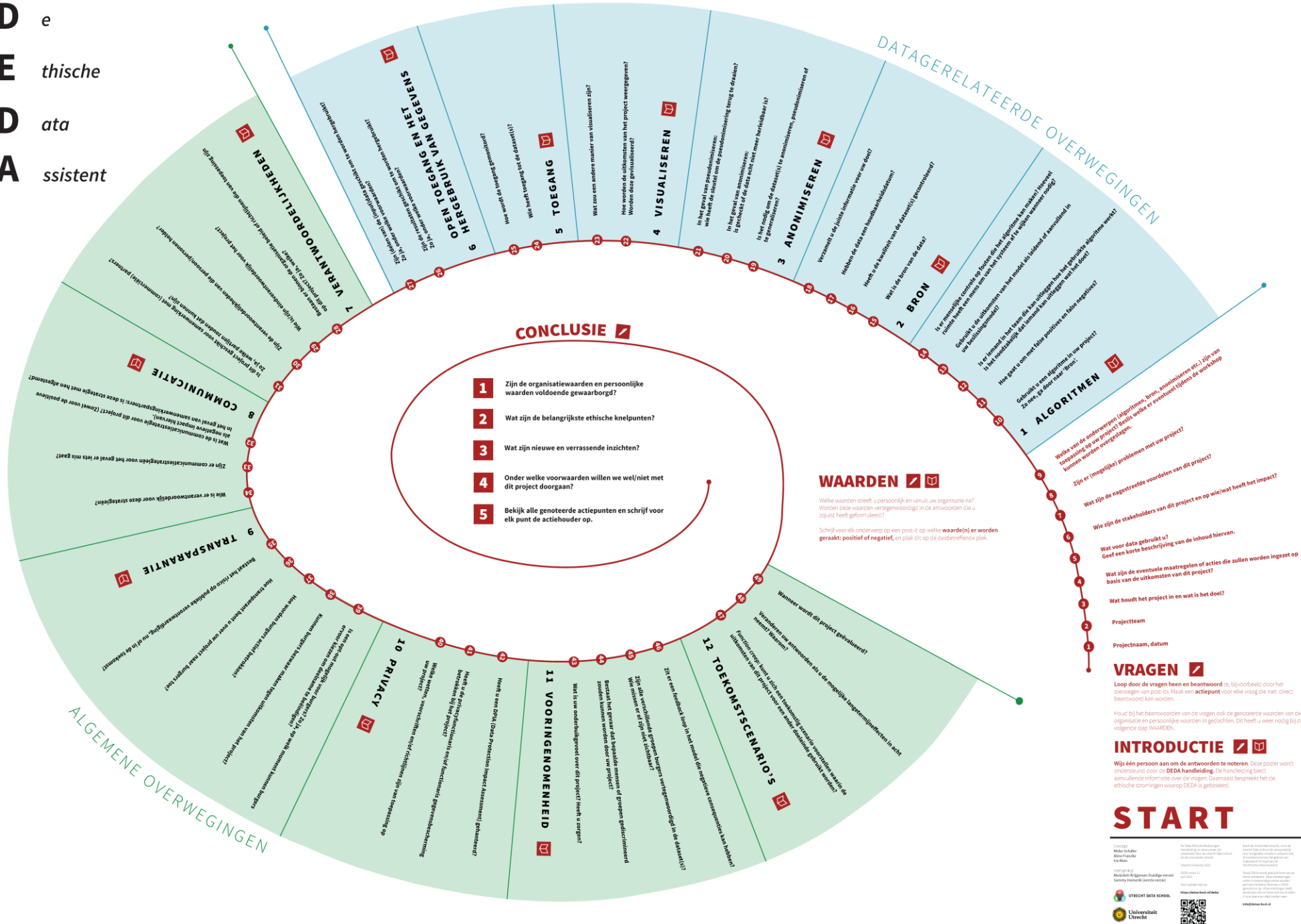


# DataWijzer

- Tool voor raadsleden
- Speciaal voor op je tablet
- Toont kritische vragen en wat je van een goed antwoord mag verwachten
- Volgt de raadscyclus:
  - **Beeldvorming**
  - **Oordeelsvorming**
  - **Besluitvorming**



# De Ethische Data Assistent



## CONCLUSIE

- 1 Zijn de organisatiewaarden en persoonlijke waarden voldoende gewaarborgd?
- 2 Wat zijn de belangrijkste ethische knelpunten?
- 3 Wat zijn nieuwe en verrassende inzichten?
- 4 Onder welke voorwaarden willen we wel/niet met dit project doorgaan?
- 5 Bekijk alle gegenoteerde actiepunten en schrijf voor elk punt de actiehouder op.

## WAARDEN

Welke waarden streeft u persoonlijk en vanuit uw organisatie na? Worden deze waarden vergeten/overgeleefd in de antwoorden die u juist heeft geformuleerd? Schrijf voor elk onderwerp op een positie op welke waarde(n) er worden geraakt: positief of negatief, en plak dit op de desbetreffende plek.

## DATAGERELATEERDE OVERWEGINGEN

- 1 **ALGORITMEN**  
Geholte e een algoritme in uw project? Hoe gaat u om met bias positief en false negatief?  
Welke van de onderwerpen (algoritmen, bron, anonimiseren, etc.) zijn van toepassing op uw project? Breek welke er eventueel tijdens de workshop kunnen worden overgeleefd?  
Zijn er (mogelijke) problemen met uw project?
- 2 **BRON**  
Waar is de bron van de data?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Verzamelde u de juiste informatie voor uw doel?  
Hoeveel data een houdbaarheidsduur?  
In het geval van pseudonimiseren: hoe heeft de bron om de pseudonimizing terug te draaien?  
In het geval van anonimiseren: is de bron van de data echt niet meer herleidbaar is?  
Is het nodig om de data(s) te anonimiseren, pseudonimiseren of te anonimiseren?
- 3 **ANONIMISEREN**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?  
In het geval van pseudonimiseren: hoe heeft de bron om de pseudonimizing terug te draaien?  
In het geval van anonimiseren: is de bron van de data echt niet meer herleidbaar is?  
Is het nodig om de data(s) te anonimiseren, pseudonimiseren of te anonimiseren?
- 4 **VISUALISEREN**  
Hoe worden de uitkomsten van het project weergegeven? Worden deze gevisualiseerd?  
Wat zou een andere manier van visualiseren zijn?  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 5 **TOEGANG**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 6 **HERBEBRIJK VAN HET TOEGANG EN HET TOEGANG**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 7 **VERANTWOORDELIJKHEID**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 8 **COMMUNICATIE**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 9 **TRANSPARANTIE**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 10 **PRIVACY**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 11 **VOORINGENOMENHEID**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?
- 12 **TOEKOMSTSCENARIO'S**  
Hoeveel data een houdbaarheidsduur?  
Verzamelde u de juiste informatie voor uw doel?  
Heeft u de kwaliteit van de data(s) gecontroleerd?  
Waar is de bron van de data?

## VRAGEN

Loop door de vragen heen en beantwoord ze, bijvoorbeeld door het toevoegen van positie. Maak een actiepunt voor elke vraag die niet direct beantwoord kan worden.

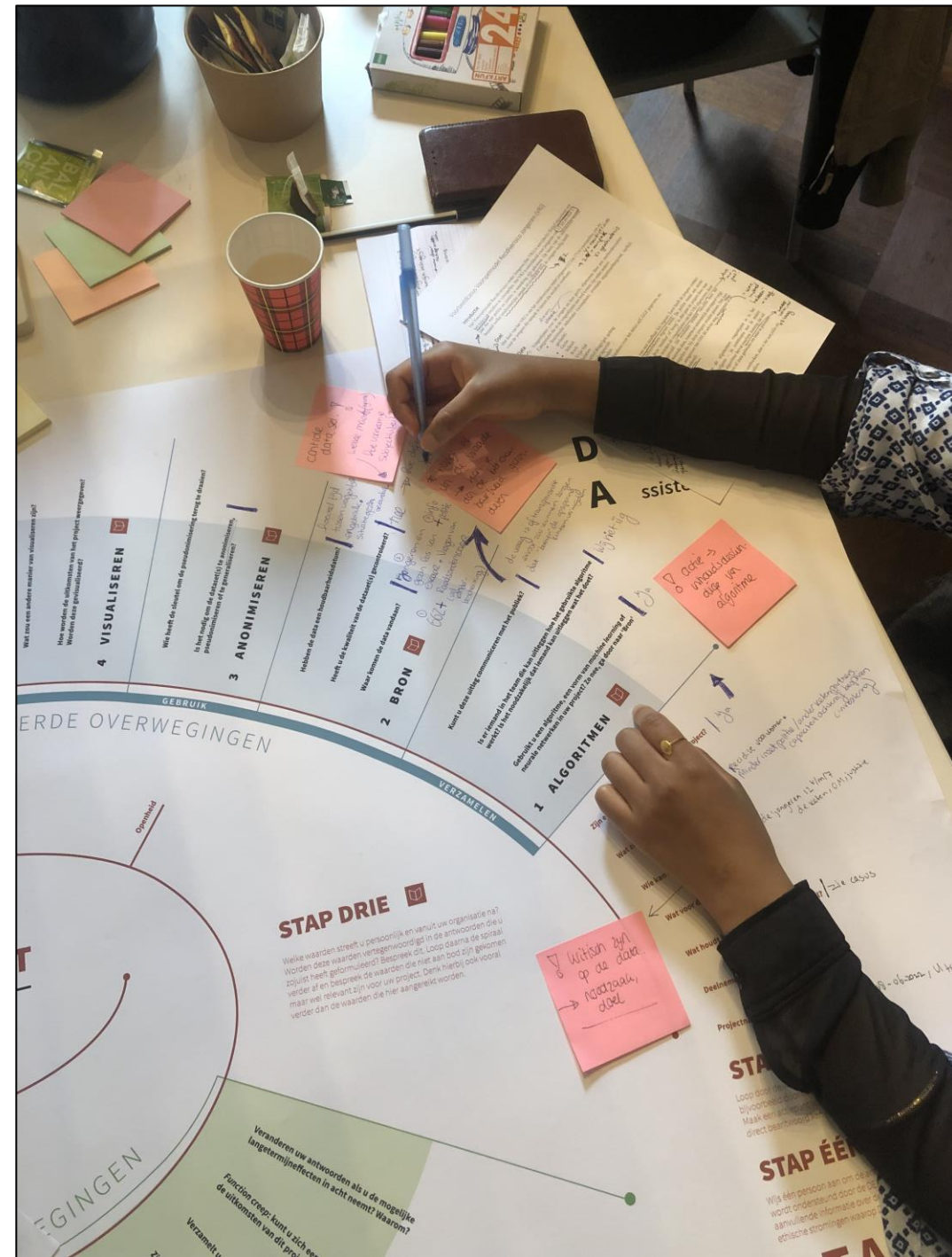
Houd bij het beantwoorden van de vragen ook de gegenoteerde waarden van de organisatie en persoonlijke waarden in gedachten. Dit heeft u weer nodig bij de volgende stap: WAARDEN.

## INTRODUCTIE

Wij's één persoon aan om de antwoorden te noteren. Deze poster wordt ontwikkeld voor de OEDA handreiking. De handreiking biedt aanvullende informatie over de vragen. Daarnaast bespreekt het de ethische stromingen waarop OEDA is gebaseerd.

# START









**DEDA-trainer**

**Communicatie adviseur**

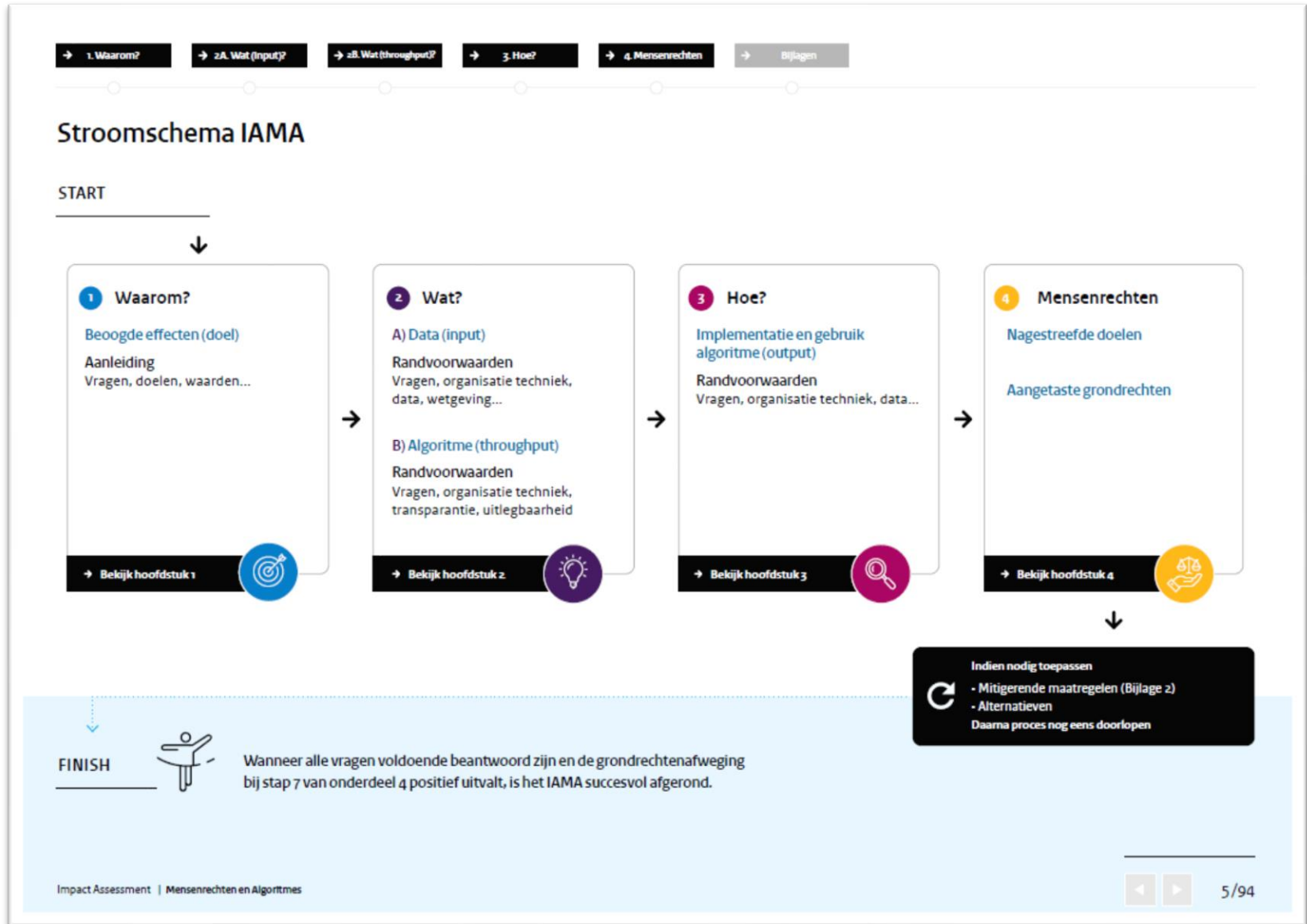
**Projectleider**

**Data-analist**

**Domein-expert**

**FG /  
privacy officer**

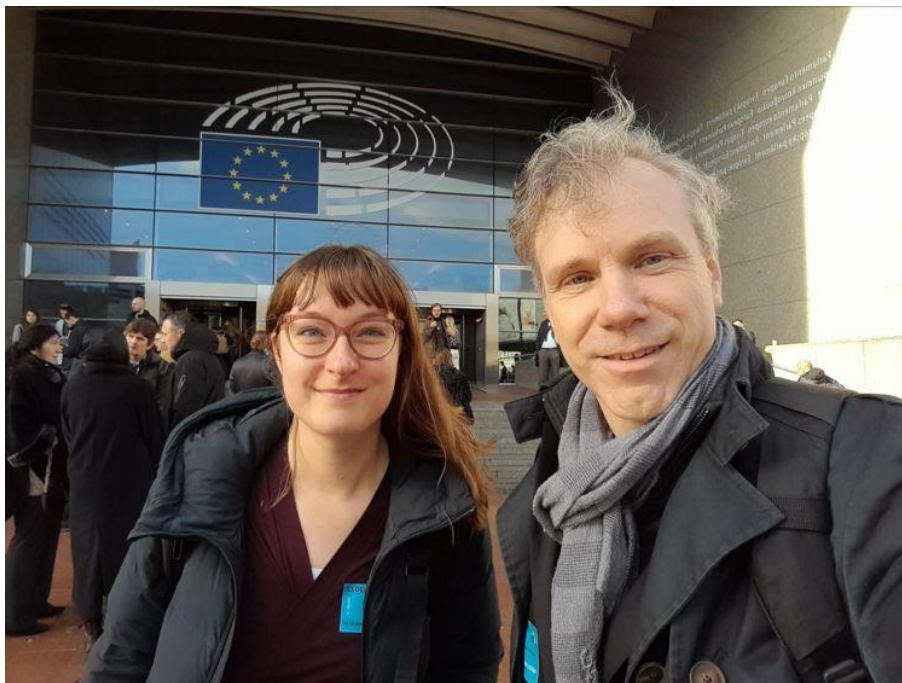
**Beleidsmedewerker**





## Tweede Kamer stemt in met verplicht gebruik Impact Assessment Mensenrechten en Algoritmes

- Moties Tweede Kamer
- Spreken op Europees Parlement
- Mogelijke toekomstige verplichting





## Bronnen

- De Ethische Data Assistent (DEDA) – poster: <https://dataschool.nl/deda-2/poster/>
- De Ethische Data Assistent (DEDA) - Remote versie: <https://dataschool.nl/deda-2/remote/>
- DataWijzer: <https://dataschool.nl/onderzoek/verantwoorde-ai/datawijzer/>
- IAMA: <https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes>

## Aanbevolen materiaal

- [Rapportage Algoritmerisico's in Nederland](#) door de Autoriteit Persoonsgegevens, juli 2023.
- Informatie over mensenrechten en digitalisering van het College voor de Rechten van de Mens: <https://www.mensenrechten.nl/themas/digitalisering>
- Informatie van de VNG over digitalisering en gemeenten: <https://vng.nl/rubrieken/informatiesamenleving>
- Informatie van Amnesty International over technologie en mensenrechten: <https://www.amnesty.nl/wat-we-doen/tech-en-mensenrechten>
- Wanneer een DPIA en wanneer een IAMA? (handreiking): <https://www.cip-overheid.nl/media/av0dmahv/20230614-gezamenlijk-gebruik-iama-en-model-dpia-rijksdienst-v1-0.pdf>



- GEM op Common Ground
  - <https://commonground.nl/groups/view/978adf62-97f2-4955-87cc-7fee0cd238e0/team-virtuele-assistent-gem>
- Open Gemeente Initiatief – Chatbot (GEM)
  - <https://opengem.nl/producten/chatbot/>
- Toets GEM aan bouwstenen Omnichannel aanpak
  - <https://vng.nl/sites/default/files/2022-03/gem.pdf>