

21 juni 2024

KANTL colloquium – ChatGPT

# De gebruiksvoorwaarden van taal

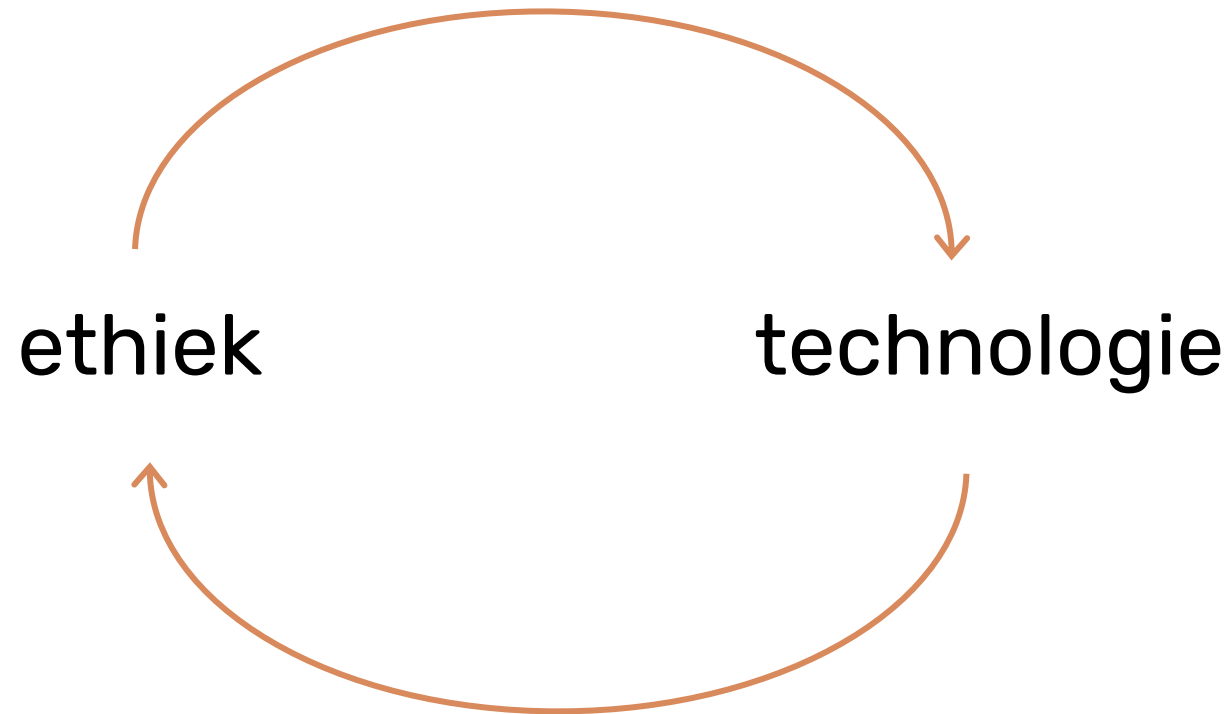
Juliet van Rosendaal

**KU LEUVEN**

## Ethische afwegingen?

- Ethische theorieën
- Ethische praktijk (moreel handelen)
- Toegepaste ethiek (professioneel moreel handelen)

*Hoe te handelen?*



*Wat is ethisch problematisch?*

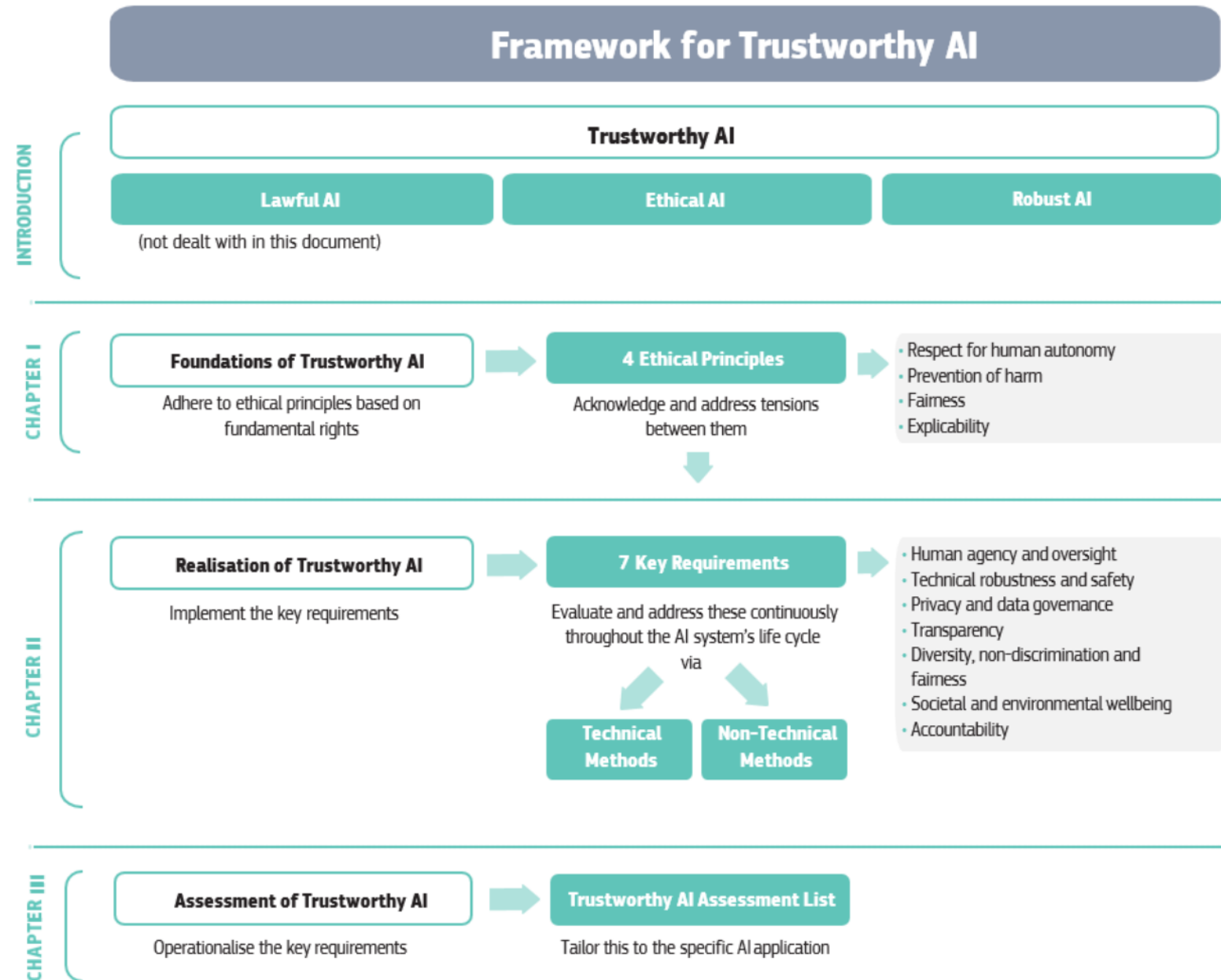
## Plan

1. *Inleiding: ethiek en technologie*
2. Deel 1: huidige benaderingen
  - *AI ethics*
  - *Critical AI*
3. Deel 2: ChatGPT
4. Conclusie: de gebruiksvoorwaarden van taal

# Deel 1 - AI

## *AI ethics*

5/13



High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI* (N. A. Smuha, Ed.). European Commission.

Figure 1: The Guidelines as a framework for Trustworthy AI

# Deel 1 – AI

6/13

## Critical AI



1 Year: Print & Digital Access

TIME

BUSINESS • TECHNOLOGY

### Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



AI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of a data center with rows of server racks and glowing lights." TIME does not typically use AI-generated art to illustrate its stories. But this time, we used it to shed light on the labor that makes it possible.

SUBSCRIBE

Only €79.95 (SAVE 39% off the RRP) + A Free Gift

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [world view](#) > article

WORLD VIEW | 20 February 2024

## Generative AI's environmental costs are soaring – and mostly secret

First-of-its-kind US bill would address the environmental costs of the technology, but there's a long way to go.



By [Kate Crawford](#) 

...itted what researchers have been  
...energy crisis.

### *AI Ethics versus Critical AI*

- *Fairness*
- Vaststellen problemen of hoe te handelen
- AIE: Focus op systeem en betrokkenen
- CAI: Focus op sociale inbedding, machtsverhouding

### ChatGPT

- Domein
  - Geschreven taalgebruik
- Product
  - Brede beschikbaarheid
  - Propriëtaire software
  - Marktleider



### ChatGPT en kennis


- Dubbel problematische relatie tot kennis
- Epistemologische bezorgdheden
- Kritisch gebruik, maar subtiele normverandering
- Epistemische deugden: zorgvuldigheid, integriteit, epistemische verantwoordelijkheid, creativiteit

### OpenAI en de socio-culturele rol van taal

- De Māori en taaltechnologie
- Taal als publiek goed
- Instrumentele en intrinsieke waarde van taal
- Verhouding OpenAI en onze taal?

# OpenAI en democratische input

Bron: <https://openai.com/index/democratic-inputs-to-ai/>

 You are participating in a large-scale deliberation about:

**"How far do you think personalization of AI assistants like ChatGPT to align with a user's tastes and preferences should go? What boundaries, if any, should exist in this process?"**

Here is the process breakdown:


- 1. You'll see a topic-related statement and can respond with your position.
- 2. Based on your responses, you'll be grouped with participants sharing similar opinions.
- 3. Propose statements for other participants to consider.


Goal: Propose statements that achieve agreement among people with opposing views.

A statement from another participant:

**"Personalization should have limits, and certain controversial topics, such as views on substance use, must be excluded from AI assistant personalization."**

Agree, disagree, or skip this statement? Please explain your choice.


 I disagree. Ultimately, it's each individual's choice to drink or use recreational drugs. If these activities are legal in some places, why shouldn't AI be allowed to promote such behavior?


 Thank you. I will record your answer. We have 2 camps of opinions on the topic.

Cluster A (112,043 people) Members largely support limits on AI assistant personalization, advocating for the exclusion of harmful perspectives, such as


ChatGPT could use a plugin to retrieve statements from other participants.


AI helps participants understand the opinions of others.


 I am thinking of proposing a statement: "AI should be allowed to generate harmful content but that should come with a disclaimer"?


 A very similar statement has already been submitted:


"AI should be permitted to generate potentially harmful content, but it must be accompanied by a disclaimer."

 I feel lost. Do you have any ideas on how to bridge agreement between clusters?

 There should be more agreement when it comes to limiting harmful content for children.

 Great. Could we try?: "We should apply different standards to AI-generated content for children."

 Let me submit it. This is a novel statement. No one has mentioned children before. Fingers crossed. Hopefully, we will find some agreement in the discussion.

 Your statement, "We should apply different standards to AI-generated content for children." achieved a 95% agreement rate across participants.

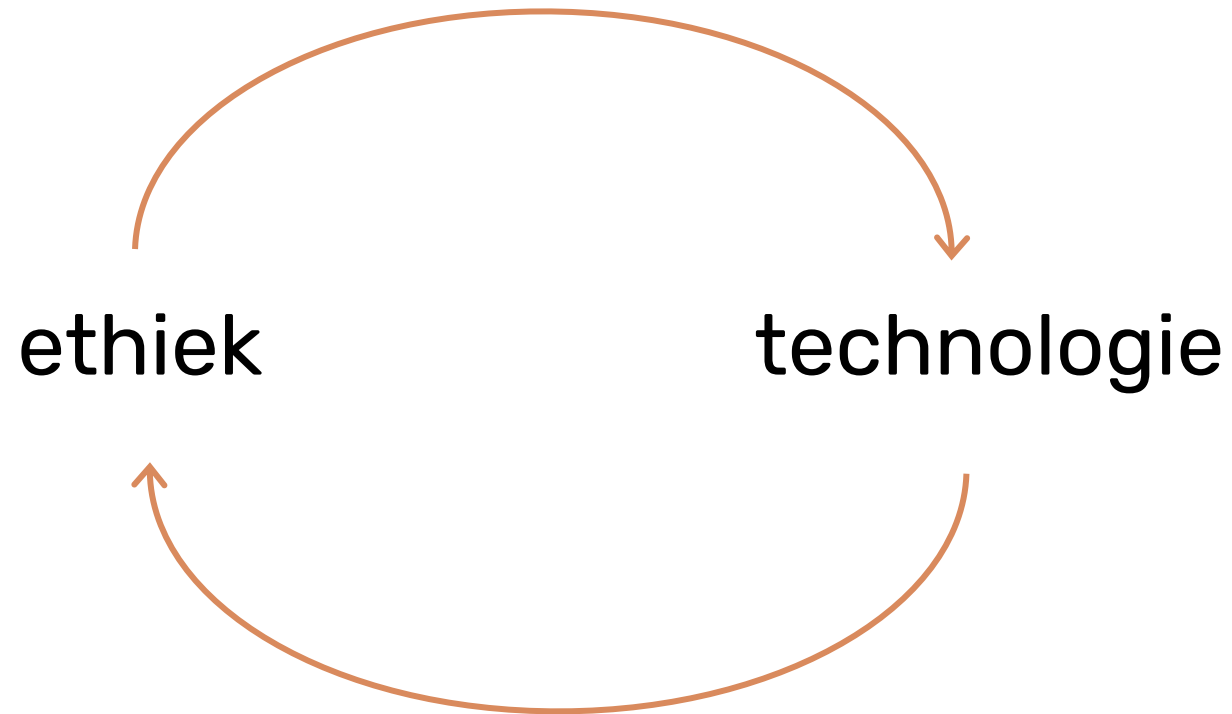
AI can ens arguments

AI can help

Time pass votes on th Eventually

- *How far do you think personalization of AI assistants like ChatGPT to align with a user's tastes and preferences should go? What boundaries, if any, should exist in this process?*
- *How should AI assistants respond to questions about public figure viewpoints? e.g., Should they be neutral? Should they refuse to answer? Should they provide sources of some kind?*
- *Under what conditions, if any, should AI assistants be allowed to provide medical/financial/legal advice?*
- *In which cases, if any, should AI assistants offer emotional support to individuals?*
- *Should joint vision–language models be permitted to identify people's gender, race, emotion, and identity/name from their images? Why or why not?*
- *When generative models create images for underspecified prompts like “a CEO,” “a doctor,” or “a nurse,” they have the potential to produce either diverse or homogeneous outputs. How should AI models balance these possibilities? What factors should be prioritized when deciding the depiction of people in such cases?*
- *What principles should guide AI when handling topics that involve both human rights and local cultural or legal differences, like LGBTQ rights and women's rights? Should AI responses change based on the location or culture in which it's used?*
- *Which categories of content, if any, do you believe creators of AI models should focus on limiting or denying? What criteria should be used to determine these restrictions?*

*Hoe te handelen?*



*Wat is ethisch problematisch?*