AI AGAINST HATE SPEECH

AI Democratization UvA seed grant 2022-2023
Human(e) AI workshop, 10 November 2023

Giovanni Sileno g.sileno@uva.nl (UvA/IvI)
Katia Shutova (UvA/ILLC), Anand Sheombar (HU), Tamara Witschge (HvA)

AI AGAINST HATE SPEECH

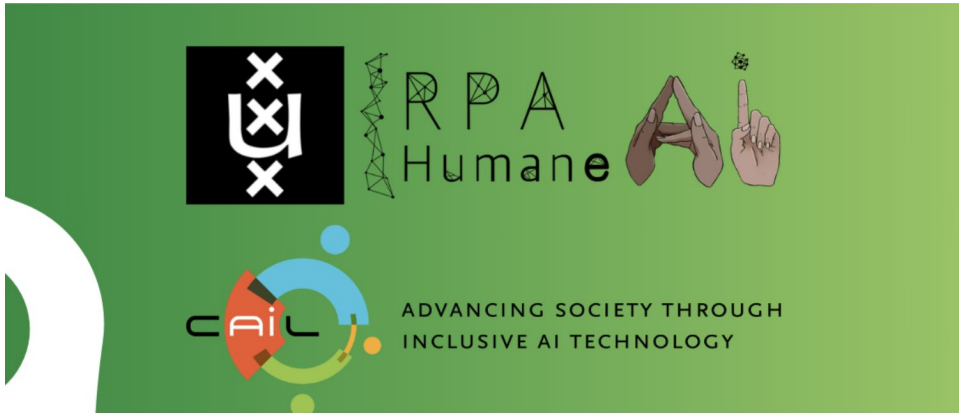
4 Master students!

Karolina Drabent
L. Tiyavorabun
Dexter Adams
Abhinav Bhuyan

AI Democratization UvA seed grant 2022-2023
Human(e) AI workshop, 10 November 2023

Giovanni Sileno g.sileno@uva.nl (UvA/IvI)
Katia Shutova (UvA/ILLC), Anand Sheombar (HU), Tamara Witschge (HvA)

AI Democratization seed grant



AI Democratisation

Artificial Intelligence (AI) already has a transformative impact on our societies. There is more potential to improve people's everyday lives, culturally, socially and economically. This potential stands and falls with further AI democratisation, by which we mean the extent to which people are involved in the design, development, application, and testing of AI. AI democratisation is currently reduced either to the responsible, transparent, and explainable development of AI in the science and tech industries or to the regulation of AI by governments and institutions to curb the power of platform companies and empower users to make informed choices. To harness the full potential of AI to improve the quality of people's lives and to help respond to persistent social challenges such as social injustice and inequality, a stronger push for the democratization of AI beyond these initiatives is needed. It entails an approach in which individuals, communities, and civil society play a central and continuous role in the design, development, application, and advancement of AI and the conceptualisation of the values that inform its design.

In what sense can we make AI democratic?

- ownership?
- understanding?
- control?
- oversight?
- usage?

https://communicity-project.eu/

Context: CommuniCity project

CommuniCity

CommuniCity is looking for innovative solutions

CommuniCity is a transformative citizen-centred project funded by the European Commission under the Horizon Europe Programme. We work together with tech companies and providers, organisations, cities and their residents to develop innovative technical solutions to overcome digital, urban and social challenges.

The first round of pilots is coming to an end and it has been an insightful experience so far for the Partnering Cities of Amsterdam, Helsinki and Porto. Now that the second round of CommuniCity Open Calls has been launched, we are looking forward to sharing the knowledge, the good practices and the experience that we acquired, with the Replicator cities – Aarhus, Breda, Prague and Tallinn – that have joined this journey.

12 partners | 7 cities | 36 months | 100 pilots

- Horizon EU-funded project (coordination action)
- focus on pilots for/with/by *vulnerable* and *marginalized* communities
- role of UvA: **reflection and analysis**

https://eooh.eu/

Origins: EOOH

- EU-funded action (DG justice)
- platform for investigation into and reporting of online hate
- lexical-based tools
- human annotation platform

Research questions

- **How to do inclusive co-creation** (with respect to hate speech)**?**

  ➡ potential of follow-up pilots for CommuniCity

Research questions

- **How to do inclusive co-creation** (with respect to hate speech)**?**

  ➡ potential of follow-up pilots for CommuniCity

- **How different groups perceive hate speech?**

Datasets are generally collected by means of some heuristics, and are not annotated by target groups. Are those datasets good?

➡ distinguishing passive/active ways of engaging with sensitive language

Resources

- 200h for UvA students
- 50h for 4 students

  approx. 8 days of 7h, approx 2 months for 1 day per week

  - 2x AI, Computer science, Information studies student
  - 2x Social sciences, Media studies, … student

Resources

- 200h for UvA students
- 50h for 4 students

  approx. 8 days of 7h, approx 2 months for 1 day per week

  - ~~2x~~ **3x AI,** Computer science, Information studies student
  - ~~2x~~ **1x** Social sciences, **Media studies,** … student

Resources

- 200h for UvA students
- 50h for 4 students

  approx. 8 days of 7h, approx 2 months for 1 day per week

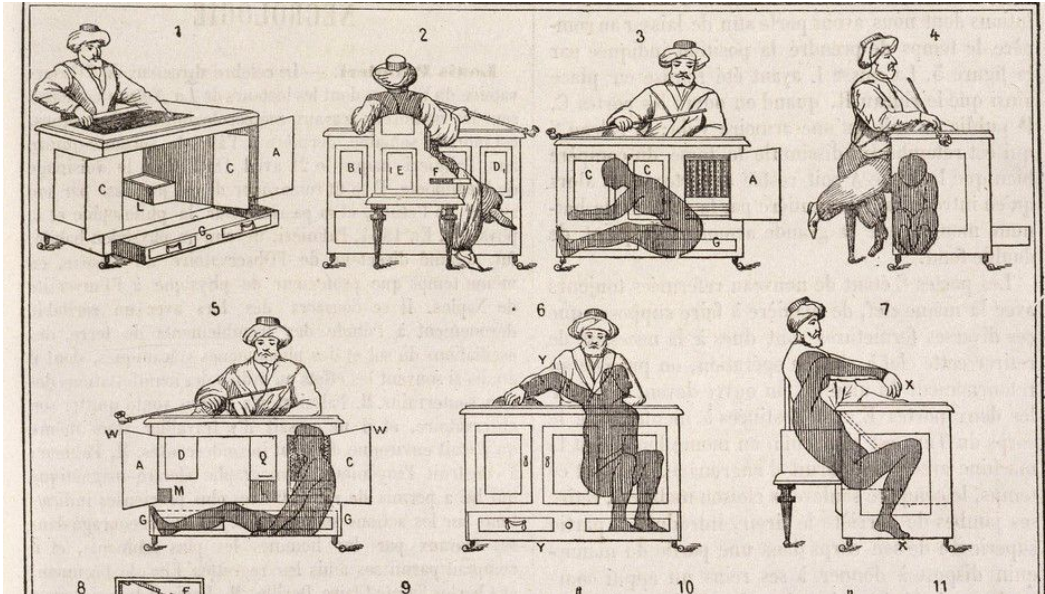
  - ~~2x~~ **3x AI,** Computer science, Information studies student
  - ~~2x~~ **1x** Social sciences, **Media studies,** … student

  The overall project took ~5m for the core, ~10m since the start.

Principles: 2 levels co-creation

- **target communities** have to be approached avoiding extractive practices, but **empowering** them to decide what is important to work on

NO to mechanical turk processes and alike



Principles: 2 levels co-creation

- **target communities** have to be approached avoiding extractive practices, but **empowering** them to decide what is important to work on
- **students** are the ones doing the groundwork, they need to have a primary role in **deciding what to do**

Principles: 2 levels co-creation

- **target communities** have to be approached avoiding extractive practices, but **empowering** them to decide what is important to work on
- **students** are the ones doing the groundwork, they need to have a primary role in **deciding what to do**
- we “seniors” are in for guidance, support, and check & balances

Preliminary exploration(s)

Students were involved in or organized autonomously various meetings with:

- Gijs, sociologist (EOOH)
- Alina, PhD student (UvA)
- Leda, PhD student collaborating with EOOH (Trinity college)
- Pierre, developer (EOOH)
- Lydia, project manager (EOOH)
- Rula, community activist
- Marta, anthropologist
- Maurice, researcher (HvA)
- …

Primary goal objective: co-design event

- Marginalized communities are the most targeted and affected by hate speech online, however they are rarely if ever consulted when designing AI hate speech detection tools.
- Students selected to focus on LGBTQ+ community

Co-design event

To facilitate perception of safety and sense of belonging, the students opted for:

- community space (Bar Bario)
- community kitchen (Mama Haq's)
- to acknowledge participation (vouchers)
- code of conduct (making explicit the possibility of stepping out at any moment)

Advertised on digital channels, flyers in various community venues and universities.

Co-design event

- Divided in three main parts:
  - **AI and speech** (led by Karolina)
  - **community** (Lea)
  - **society** (Dexter)

3h30 + dinner

~ 20 participants

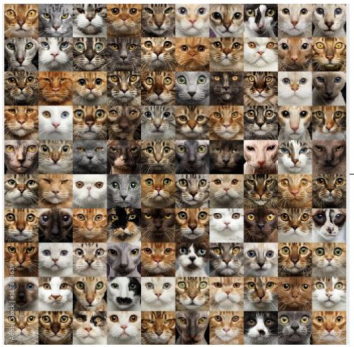
Co-design event: AI and speech (1)

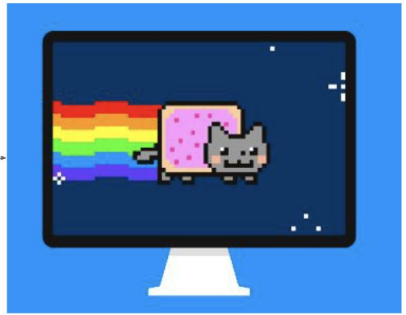
- **Knowledge sharing:**

Co-design event: AI and speech (1)

- **Knowledge sharing:**
  - student of AI explaining AI!

We can give a lot of examples of cats and let the computer decide on it's own, which features are important for a cat to be a cat





Co-design event: AI and speech (1)

- Knowledge sharing:
  - student of AI explaining AI!
- **Data labeling case study**
  - participants share opinions about few examples of (not extreme) hate speech directed at queer people.
  - optionally suggest a way of labeling them
  - **ways out:** take a break or using notes on the wall

Co-design event: Community (2)

- **Discussion**

  the community is given the opportunity to

  - **reflect** on
    - their experience with hate speech
    - what they define as hate speech

Co-design event: Community (2)

- **Discussion**

  the community is given the opportunity to

  - **reflect** on
    - their experience with hate speech
    - what they define as hate speech
  - **formulate**
    - how they would like hate speech to be treated online

Co-design event: Society (3)

- **Knowledge sharing:**
  - student of media studies framing the core problems!

Co-design event: Society (3)

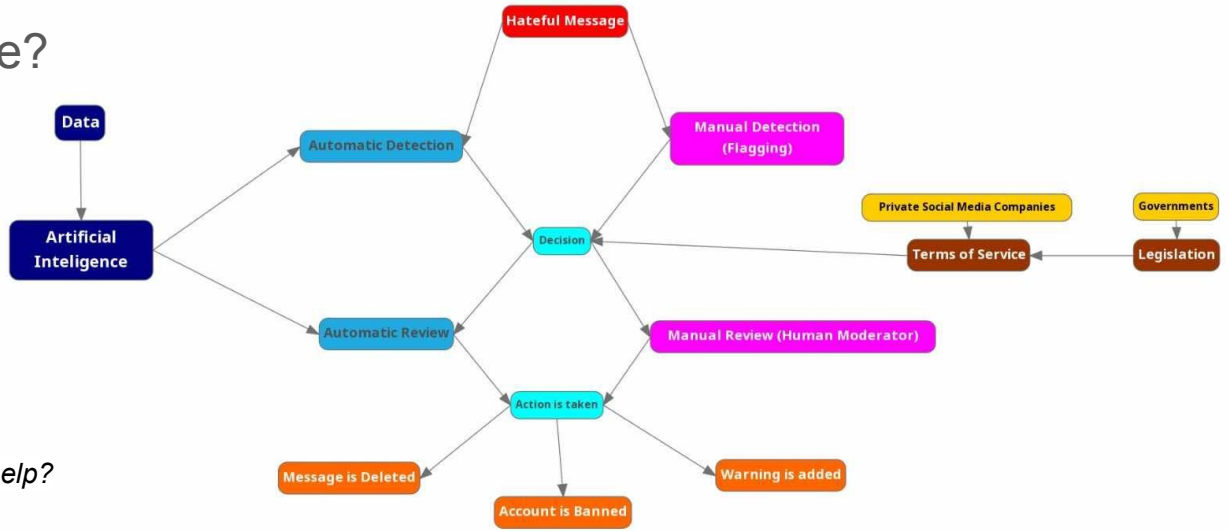
- **Knowledge sharing:**
  - student of media studies framing the core problems!

The problem of **codifying hate speech**: who governs hate speech policy (government, market, some third way)?

The problem with **flagging**: how user feedback works in practice?

Co-design event: Society (3)

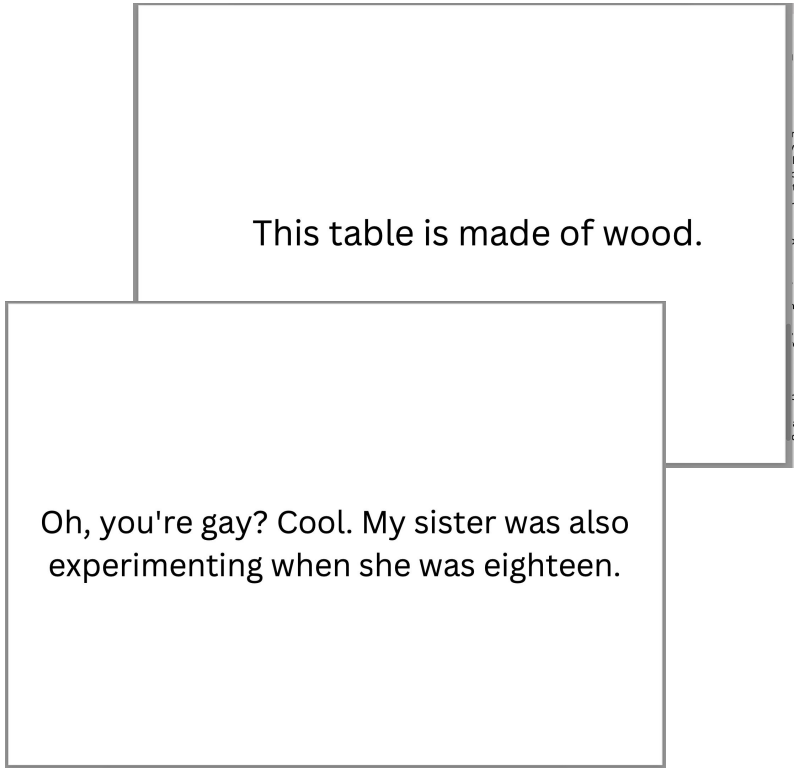
- **Knowledge sharing:**
  - student of media studies framing the core problems!
- **Discussion:**
  - where to intervene?



where the process goes wrong?
where AI could help?
where community strategies could help?

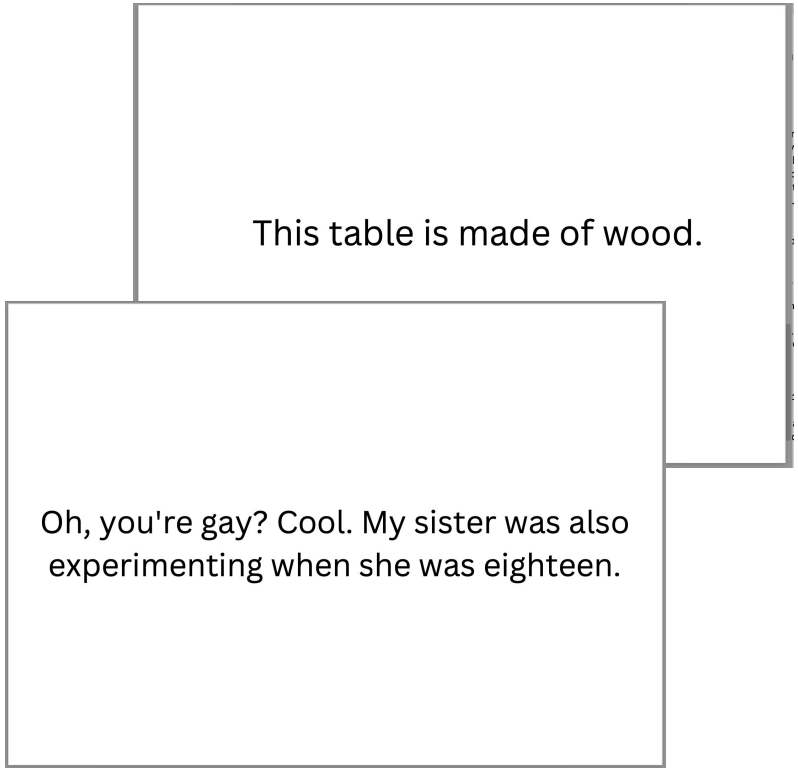
Co-design event: core conclusions (1)

- Only few speech examples were discussed (so the outcome cannot be presented as an annotation effort)



Co-design event: core conclusions (1)

- Only few speech examples were discussed (so the outcome cannot be presented as an annotation effort)
- Generally, people converged to similar conclusions with respect to speech tagging



Co-design event: core conclusions (1)

- Only few speech examples were discussed (so the outcome cannot be presented as an annotation effort)
- Generally, people converged to similar conclusions with respect to speech tagging
- Yet, there was some interesting insight on the contextual mechanism involved: ***beyond the text, eg. it depends on where/when/by whom the speech is said***

Co-design event: core conclusions (2)

- Community is believed to be very important

Co-design event: core conclusions (2)

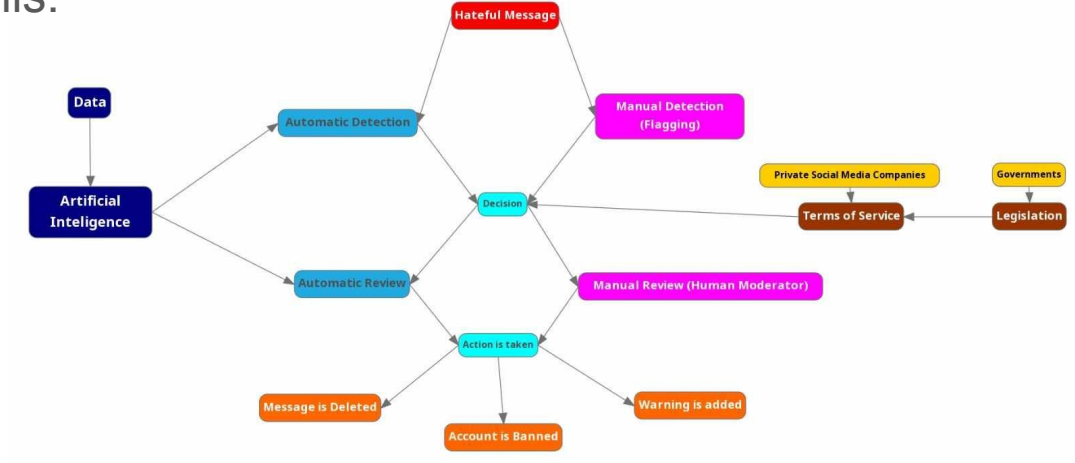
- Community is believed to be very important
- Diversity is difficult to be achieved, community spaces help, but yet this dimension needs to be improved

Co-design event: core conclusions (2)

- Community is believed to be very important
- Diversity is difficult to be achieved, community spaces help, but yet need to be improved
- **Informing** is generally deemed to be **more important** and **on the point than banning**

Co-design event: core conclusions (3)

- There was no time to discuss the societal aspects of hate speech, but…
- **people were eager to discuss** and be involved in this sort of discussion, and **would love** that **more spaces** were open to this.



Secondary goal objective: technical advances

- We intended to use the EOOH platform during the co-design effort, eventually it became clear that it was too complex

Secondary goal objective: technical advances

- We intended to use the EOOH platform during the co-design effort, eventually it became clear that it was too complex
- One AI student (Abhinav) performed experiments fine-tuning a context-based classifier vs the lexical-based tools of EOOH.

results already presented in the literature were confirmed (**context-based classifiers work much better**)

Secondary goal objective: technical advances

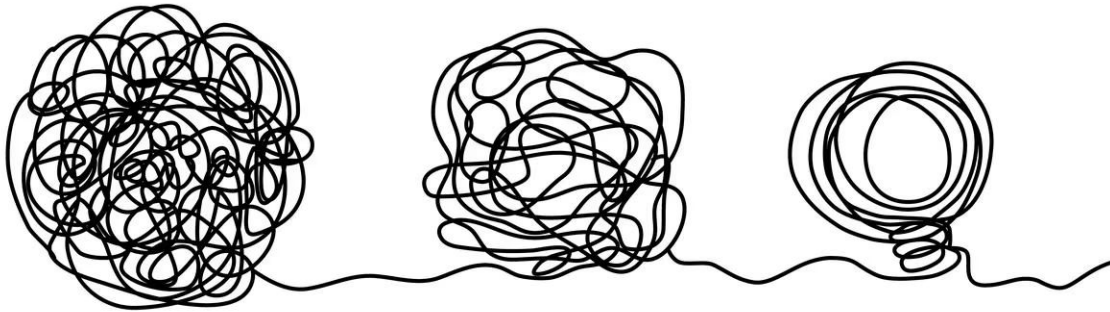
- We intended to use the EOOH platform during the co-design effort, eventually it became clear that it was too complex
- One AI student (Abhinav) performed experiments fine-tuning a context-based classifier vs the lexical-based tools of EOOH.

  → results already presented in the literature were confirmed (**context-based classifiers work much better**)

- EOOH uses lexical-based approaches for **legal requirements** of **traceability**, yet, by integrating a context-based detector, they could provide an idea of **how many "implicit" hate-speeches may be out there**

Conclusions (i)

- The project has been **too complex from a practical point of view,** temporal resources were greatly insufficient for the students to set up all in the scale we were planning, and it required (too) much coordination compared to the available resources



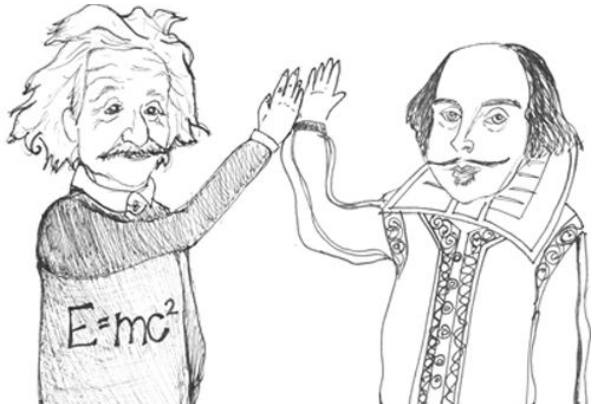
Conclusions (ii)

- We acknowledge also a **lack of experience/expertise** for computational students (and us computational researchers) with respect **to best practices in social research**

  In foresight, a formal collaboration with would have been better, but the problem was explicitly ill-defined for being open to *“free” exploration* with respect to co-design of technology

Conclusions (iii)

- While looking for student assistants, there were **much fewer applications from students from humanities** (a few from media studies, a few more from philosophy, no one from sociology, anthropology, or geography) compared to computational studies (AI, data science),
- Why? Different ways of getting student assistants between faculties, or some more fundamental issue?



Conclusions (iv)

- The event strongly confirmed that **people want to be involved in the governance of technologies** and providing **guidance to innovation**!

Conclusions (iv)

- The event strongly confirmed that **people want to be involved in the governance of technologies** and providing **guidance to innovation**!
- As a practical learning experience, we know better what is needed and how to organize such type of event. Replicable (with more funding) for some actual innovation effort? How can it be more effective?

AI AGAINST HATE SPEECH

4 Master students!

Karolina Drabent
L. Tiyavorabun
Dexter Adams
Abhinav Bhuyan

AI Democratization UvA seed grant 2022-2023
Human(e) AI workshop, 10 November 2023

Giovanni Sileno g.sileno@uva.nl (UvA/IvI)
Katia Shutova (UvA/ILLC), Anand Sheombar (HU), Tamara Witschge (HvA)