

# Using Digital Tools to Tackle Online Harms and Extremism

23 Nov 2023

Zeeraq Talat  
[z@zeeraq.org](mailto:z@zeeraq.org) | [@zeeraqtalat](https://www.instagram.com/zeeraqtalat)  
www: [zeeraq.org](http://zeeraq.org)

# Agenda

- Quick deep dive into AI for language technology
- Identify and discuss some weaknesses of AI
- Practical methods for inclusion in your work
- Defining and finding useful errors

# Natural Language Processing (NLP)

Operates on a foundational assumption: The Distributional Hypothesis

- Semantic (i.e., literal) meaning can arise from word frequencies

# AI for Classification

1. Develop language model
2. Create training data
3. Further develop language model with training data
4. Use the resulting model to classify new data

# AI for Classification

**TIME**

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

1. Develop language model
2. Create training data
3. Further develop language model with training data
4. Use the resulting model to classify new data

# AI for Content Moderation

1. Existing models developed for general purpose
2. Training data difficult to get right

# AI for Content Moderation

1. Existing models developed for general
2. Training data difficult to get right





A person whose reflection is being distorted by mirrors.  
Source: funplanners.com

# Working with AI

1. Developing your data for your needs
2. Identifying crucial failure modes for your case
3. Attend to the kind of errors that are most consequential

# Recommendations and Conclusion

- Develop own data with experts
- Identify the particularly weaknesses of data
- Assume AI will be wrong in useful ways
  - When identifying a particular error, look for more of them