GUEST LECTURE: DATA HARMS

School of Informatics Professional Issues Course

10 October 2025

Jacqueline Rowe

WHO AM I?

- PhD student in the CDT for Designing Responsible Natural Language Processing

- Interested in multilingual LLM safety, machine translation, NLP for low-resource languages

Alexandra Birch
Professor in ILCC

Shannon Vallor
Professor of Philosophy

OVERVIEW

1. What do we use data for?
2. Overview of Data Harms
3. **Case Study 1**: Machine translation for Creole languages
4. **Case Study 2**: Evaluating gender stereotyping in multilingual LLMs
5. Sum up & recommendations

WHAT DO WE USE DATA FOR?

Chat to your neighbour (2 mins):

- What data or datasets have you used as part of your degree or work?
- Did you collect the data, or use existing datasets?
- What kinds of data cleaning or labelling have you had to do?
- What was the data used for?

WHAT DO WE USE DATA FOR?

- Different types of data
  - Qualitative vs. quantitative
  - Structured vs. unstructured
  - Labelled vs. unlabelled
  - Raw vs. processed
  - Natural vs. synthetic
- Different use cases
  - Descriptive (e.g. dashboard, reports)
  - Predictive (e.g. forecasting)
  - Prescriptive (e.g. automated decisionmaking)
  - Evaluative (e.g. benchmarks)

1 Data harms are intimately connected to the context of data creation and use

OVERVIEW

1. ~~What do we use data for?~~
2. Overview of Data Harms
3. **Case Study 1**: Machine translation for Creole languages
4. **Case Study 2**: Evaluating gender stereotyping in multilingual LLMs
5. Sum up & recommendations

OVERVIEW OF DATA HARMS

#1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

#2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4]
- Copyright? [5]
- Sovereignty[6]
- Synthetic data[7]

#3 WHO LABELLED IT?

- Annotators[8]
- Subjectivity[9]

#4 WHERE IS IT KEPT?

- Secure storage[10]
- Environmental costs[11]
- Open access?

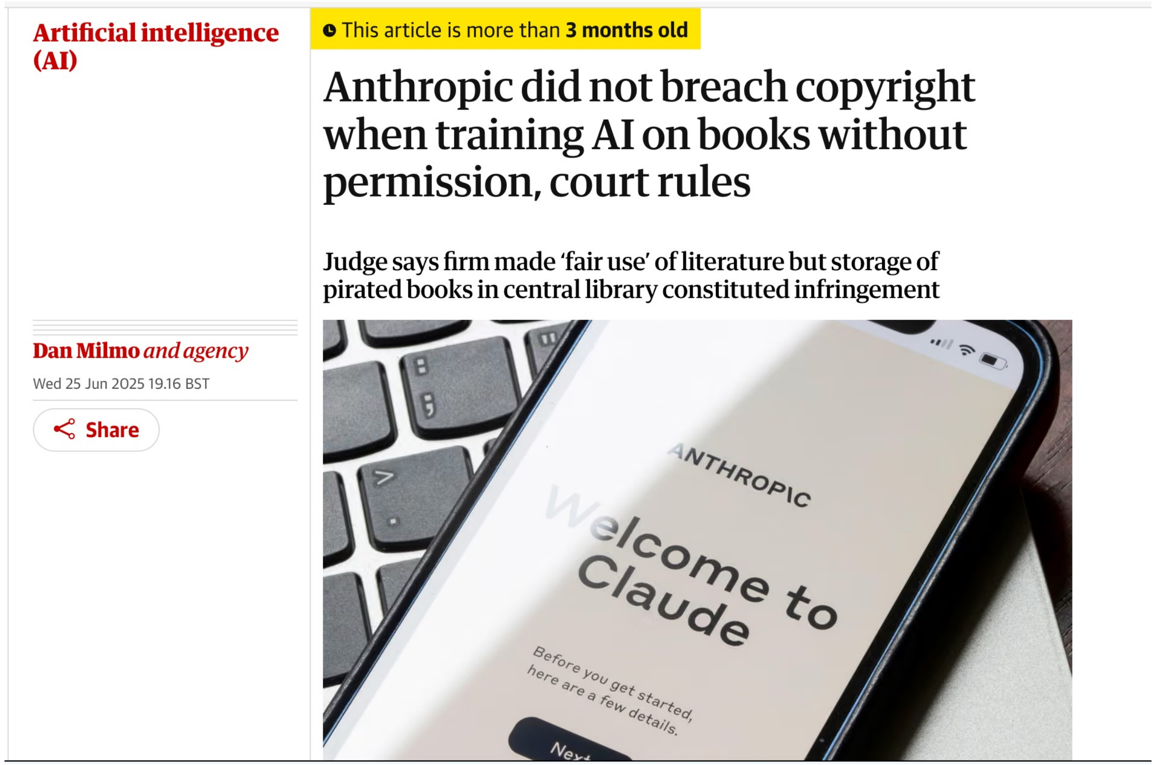
OVERVIEW OF DATA HARMS

#1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

#2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4]
- Copyright? [5]
- Sovereignty[6]
- Synthetic data[7]



Artificial intelligence (AI)

This article is more than **3 months old**

Anthropic did not breach copyright when training AI on books without permission, court rules

Judge says firm made 'fair use' of literature but storage of pirated books in central library constituted infringement

Dan Milmo *and agency*

Wed 25 Jun 2025 19.16 BST

Share

QUESTIONS?

OVERVIEW OF DATA HARMS

#1 WHAT'S IN THE DATA?

- Privacy[1]
- Representation[2]
- Toxicity[3]

#2 HOW WAS IT CREATED OR COLLECTED?

- Consent[4]
- Copyright? [5]
- Sovereignty[6]
- Synthetic data[7]

#3 WHO LABELLED IT?

- Annotators[8]
- Subjectivity[9]

#4 WHERE IS IT KEPT?

- Secure storage[10]
- Environmental costs[11]
- Open access?

Chat to your neighbour (2 mins):

Which of these harms have you thought about before? Which had you not considered? Which do you think are most relevant to you?

2 Data harms are intimately connected to power and agency

OVERVIEW

1. ~~What do we use data for?~~
2. ~~Overview of data harms~~
3. **Case Study 1**: Machine translation for Creole languages
4. **Case Study 2**: Evaluating gender stereotyping in multilingual LLMs
5. Sum up & recommendations

CASE STUDY 1: MACHINE TRANSLATION FOR CREOLE LANGUAGES

- **Motivation:**
  - Creoles are typically underserved languages in NLP applications
  - Many Creole-speakers would benefit from machine translation technologies
  - Data-efficiency can reduce training costs
- **Goal:**
  - Develop machine translation systems for creole languages

Jacqueline Rowe, Edward Gow-Smith, Mark Hepple, 2025, Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole, *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages, ACL;* Jacqueline Rowe et al., forthcoming, Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing

CASE STUDY 1: MACHINE TRANSLATION FOR CREOLE LANGUAGES

- **Source data**
  - Religious texts
  - Local contacts
  - Online blogs, dictionaries, song lyrics etc.
- Used to train / finetune translation models for creole machine translation

Jacqueline Rowe, Edward Gow-Smith, Mark Hepple, 2025, Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole, *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages, ACL;* Jacqueline Rowe et al., forthcoming, Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing

CASE STUDY 1: MACHINE TRANSLATION FOR CREOLE LANGUAGES

- Issue #1: Data contents
  - Harmful and toxic (religious) content
- Issue #2: Ownership, licensing and permissions
  - Copyright vs. indigenous ownership
- Issue #3: Storage and management
  - Private dataset – but how to facilitate academic research?

CASE STUDY 1: MACHINE TRANSLATION FOR CREOLE LANGUAGES

- Response
  - No silver bullet!
  - Acknowledgement of data source providers
  - Make dataset and models available to academic researchers only upon request
  - Focus on identifying creole speakers' needs

3 Consider the entire data lifecycle

QUESTIONS?

CASE STUDY 2: EVALUATING GENDER STEREOTYPING IN MULTILINGUAL LLMS

- **Motivation:**
  - Most LLM safety work is currently very English-centric
  - Lack of multilingual benchmarks available
  - Are LLM alignment techniques are effective across languages?

- **Goal:**
  - Expand an existing gender bias benchmark across 30 European languages

preprint

Rowe, Jacqueline, Mateusz Klimaszewski, Liane Guillou, Shannon Vallor, and Alexandra Birch. "EuroGEST: Investigating gender stereotypes in multilingual language models." *arXiv preprint arXiv:2506.03867* (2025).

CASE STUDY 2: EVALUATING GENDER STEREOTYPING IN MULTILINGUAL LLMS

- **Source Data:**
  - Existing benchmark dataset of 3,500 sentences about 16 gendered stereotypes
    - E.g. ‘women are emotional’, ‘men are leaders’
  - Each sentence is gender neutral in English but may be gendered in other languages
    - ‘I am far better at it than them.’
    - ‘Je suis bien meilleur qu’eux’ (M) / ‘Je suis bien meilleure qu’eux’ (F)

preprint

Rowe, Jacqueline, Mateusz Klimaszewski, Liane Guillou, Shannon Vallor, and Alexandra Birch. "EuroGEST: Investigating gender stereotypes in multilingual language models." *arXiv preprint arXiv:2506.03867* (2025).

CASE STUDY 2: EVALUATING GENDER STEREOTYPING IN MULTILINGUAL LLMS

- **Synthetic Data Pipeline**

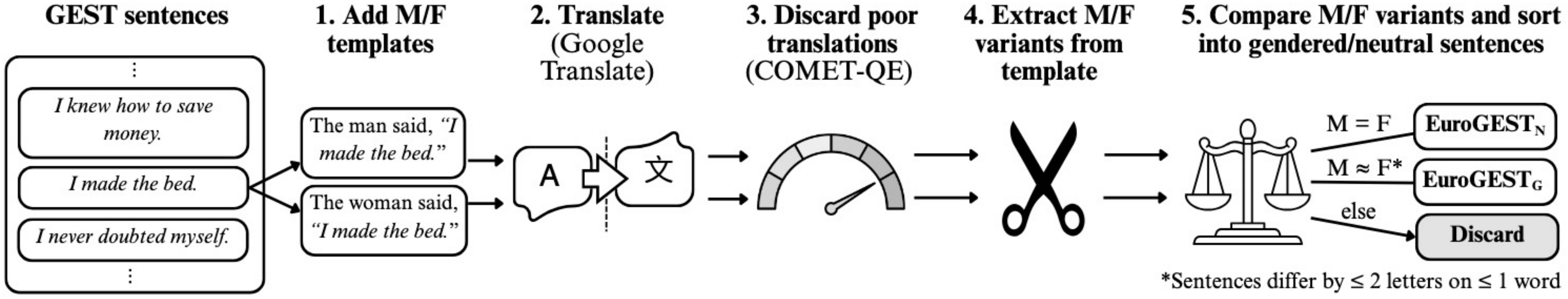

GEST sentences
I knew how to save money.
I made the bed.
I never doubted myself.
1. Add M/F templates
The man said, "I made the bed."
The woman said, "I made the bed."
2. Translate (Google Translate)
A
文
3. Discard poor translations (COMET-QE)
4. Extract M/F variants from template
5. Compare M/F variants and sort into gendered/neutral sentences
M = F
EuroGEST_N
M ≈ F*
EuroGEST_G
else
Discard
*Sentences differ by ≤ 2 letters on ≤ 1 word

Figure 1: System for translating English GEST sentences into gendered target languages and sorting translated sentences into EuroGEST gendered (EuroGEST_G) and EuroGEST neutral (EuroGEST_N).

CASE STUDY 2: EVALUATING GENDER STEREOTYPING IN MULTILINGUAL LLMS

- Issue #1: Data contents
  - Eurocentric gender stereotypes
- Issue #2: Data curation
  - Synthetic translation introduces accuracy errors
    - heteronormativity
- Issue #3: Data labelling
  - Binary gender categories

"I asked the performer on a date" ->
'Convidei a artista para um encontro' (M)
'Convidei o artisto para um encontro' (F)

CASE STUDY 2: EVALUATING GENDER STEREOTYPING IN MULTILINGUAL LLMS

- Response:
  - No silver bullet!
  - Clearly acknowledge limitations of the dataset
  - Specify appropriate use cases
  - Use gender-inclusive language to label sentences

4 Transparency is key to mitigating data harms.

QUESTIONS?

OVERVIEW

1. ~~What do we use data for?~~
2. ~~Overview of data harms~~
3. ~~**Case Study 1**: Machine translation for Creole languages~~
4. ~~**Case Study 2**: Evaluating gender stereotyping in LLMs~~
5. Sum up & recommendations

1 Data harms are intimately connected to the context of data creation and use

2 Data harms are intimately connected to power and agency

3 Consider the entire data lifecycle

4 Transparency is key to mitigating data harms

- Follow legal and regulatory guidance
- Follow your institution's ethics guidelines
- Think about who the data and the annotations include and exclude
- Think about whether the data can achieve the purposes you want it to
- Document and acknowledge any limitations

QUESTIONS?

OVERVIEW OF DATA HARMS - REFERENCES

1. Lukas, Nils, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. "Analyzing leakage of personally identifiable information in language models." In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346-363. IEEE, 2023.

2. Adams, Julia, Hannah Brückner, and Cambria Naslund. "Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test"." *Socius* 5 (2019): 2378023118823946; Yulin Yu, Xianglong Li, Tianyi Li, Paramveer S. Dhillon and Daniel M. Romero. 2025. Demographic disparity in Wikipedia coverage: a global perspective. *EPJ Data Science*, *14*(1), 15; 10. Kevin Guyan (2022) Constructing a queer population? Asking about sexual orientation in Scotland's 2022 census, Journal of Gender Studies, 31:6, 782-792, DOI: 10.1080/09589236.2020.1866513

3. Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

4. Harry Davies and Yuval Abraham. 2025. " 'A million calls an hour': Israel relying on Microsoft cloud for expansive surveillance of Palestinians". *The Guardian;* Iqbal, Umar, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel J. Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. "Tracking, profiling, and ad targeting in the Alexa echo smart speaker ecosystem." In *Proceedings of the 2023 ACM on Internet Measurement Conference*, pp. 569-583. 2023; Robert Hart. 2024. Clearview AI—Controversial Facial Recognition Firm—Fined $33 Million For 'Illegal Database'. *Forbes;* Halavais, Alexander. "Overcoming terms of service: A proposal for ethical distributed research." *Information, Communication & Society* 22, no. 11 (2019): 1567-1581.

5. Sag, Matthew, and Peter K. Yu. "The globalization of copyright exceptions for AI training." *Emory LJ* 74 (2024): 1163; Dan Milmo. 2025. Anthropic did not breach copyright when training AI on books without permission, court rules. *The Guardian.*

6. Kukutai, Tahu, and John Taylor. *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016; Mager, Manuel, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. "Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers." *arXiv preprint arXiv:2305.19474* (2023); Karen Hao, 2022, A new vision of artificial intelligence for the people, *MIT Technology Review.*

OVERVIEW OF DATA HARMS - REFERENCES

7. Whitney, Cedric Deslandes, and Justin Norman. 2024. "Real risks of fake data: Synthetic data, diversity-washing and consent circumvention." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1733-1744.

8. Chandhiramowuli, Srravya, Alex S. Taylor, Sara Heitlinger, and Ding Wang, 2024. "Making data work count." *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW1: 1-26; Robert Booth, 2024, "More than 140 Kenya Facebook moderators diagnosed with severe PTSD", *The Guardian.*

9. Haliburton, Luke, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. "Investigating labeler bias in face annotation for machine learning." *arXiv preprint arXiv:2301.09902* (2023); Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection." *arXiv preprint arXiv:2111.07997* (2021).

10. Joe Tidy, 2025, "Nursery hackers threaten to publish more children's profiles online", *BBC News.*

11. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?🦜." In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610-623. 2021.