Durham University

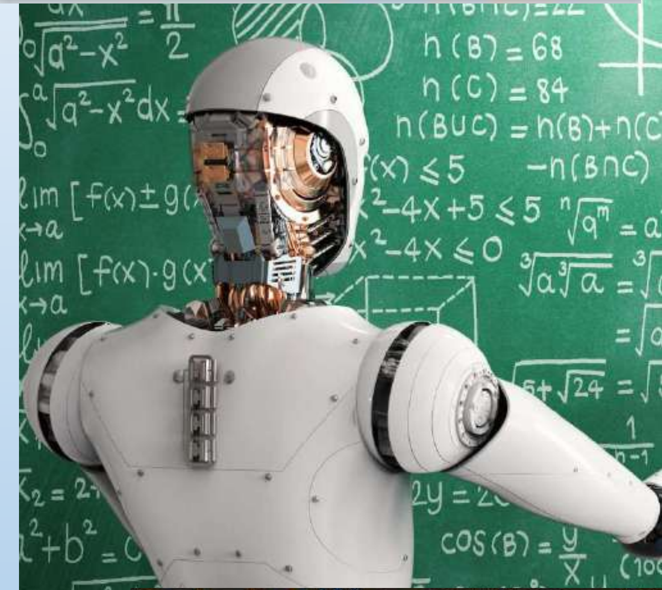
# Professor Alexandra I. Cristea

Professor of Computer Science, Durham University

# Bias in AI

# Long-term research in AIED

- AIED 2022: The 23rd International Conference on Artificial Intelligence in Education, 27-31 July, Durham University, UK AIED2022 (durham.ac.uk)
- EDM 2022 is the 15th iteration of the Educational Data Mining Conference Series EDM2022 (durham.ac.uk)
- Intelligent Tutoring Systems - 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings | Alexandra I. Cristea | Springer
- Other conferences also target this area: EC-TEL, ICALT, to name but a few



# What is bias in AI?

*(and why does it upset us)*

# What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output(‘You are right!’)
IF sees(system, my(archenemy))
THEN greet(‘You are wrong!’)
```

# What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

- 'black-box' shallow NN: train on

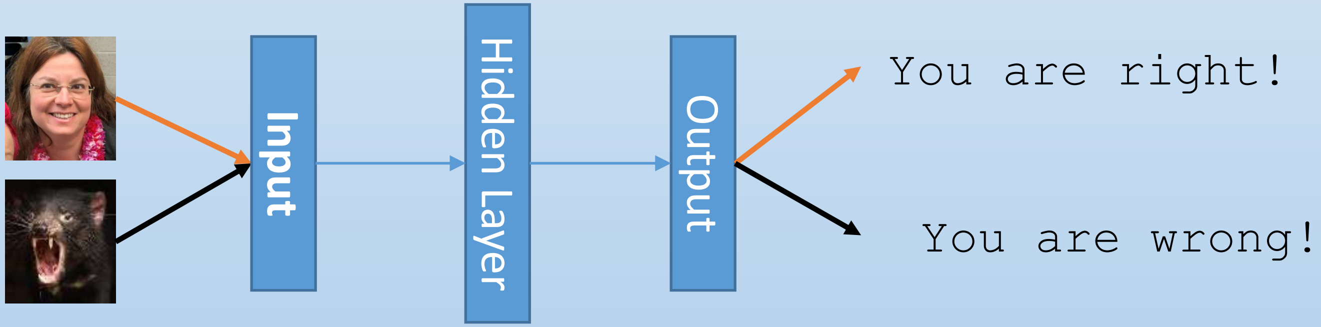
- 'black-box' deep NN:

# What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output(‘You are right!’)
IF sees(system, my(archenemy))
THEN greet(‘You are wrong!’)
```

- ‘black-box’ shallow NN: train on



```
You are right!
You are wrong!
```

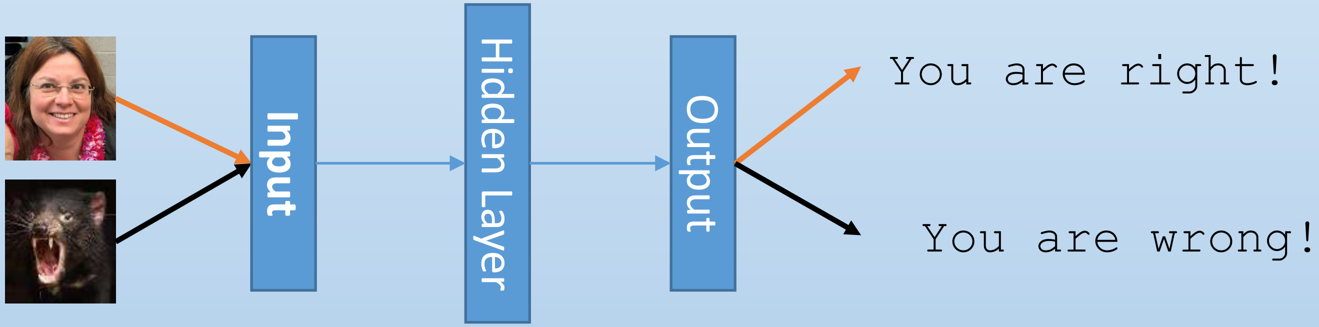
- ‘black-box’ deep NN:

# What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output(‘You are right!’)
IF sees(system, my(archenemy))
THEN greet(‘You are wrong!’)
```

- ‘black-box’ shallow NN: train on



Input

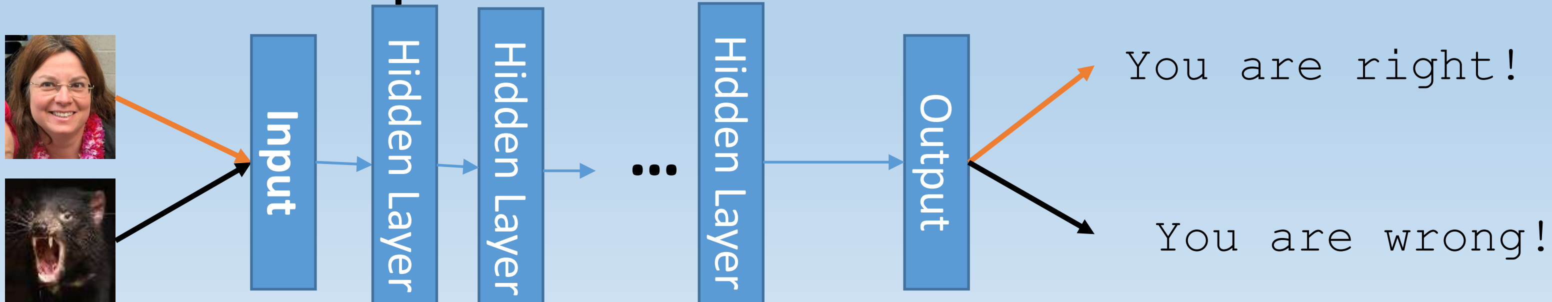
Hidden Layer

Output

You are right!

You are wrong!

- ‘black-box’ deep NN: train on



Input

Hidden Layer

Hidden Layer

...

Hidden Layer

Output

You are right!

You are wrong!

# Bias in real life

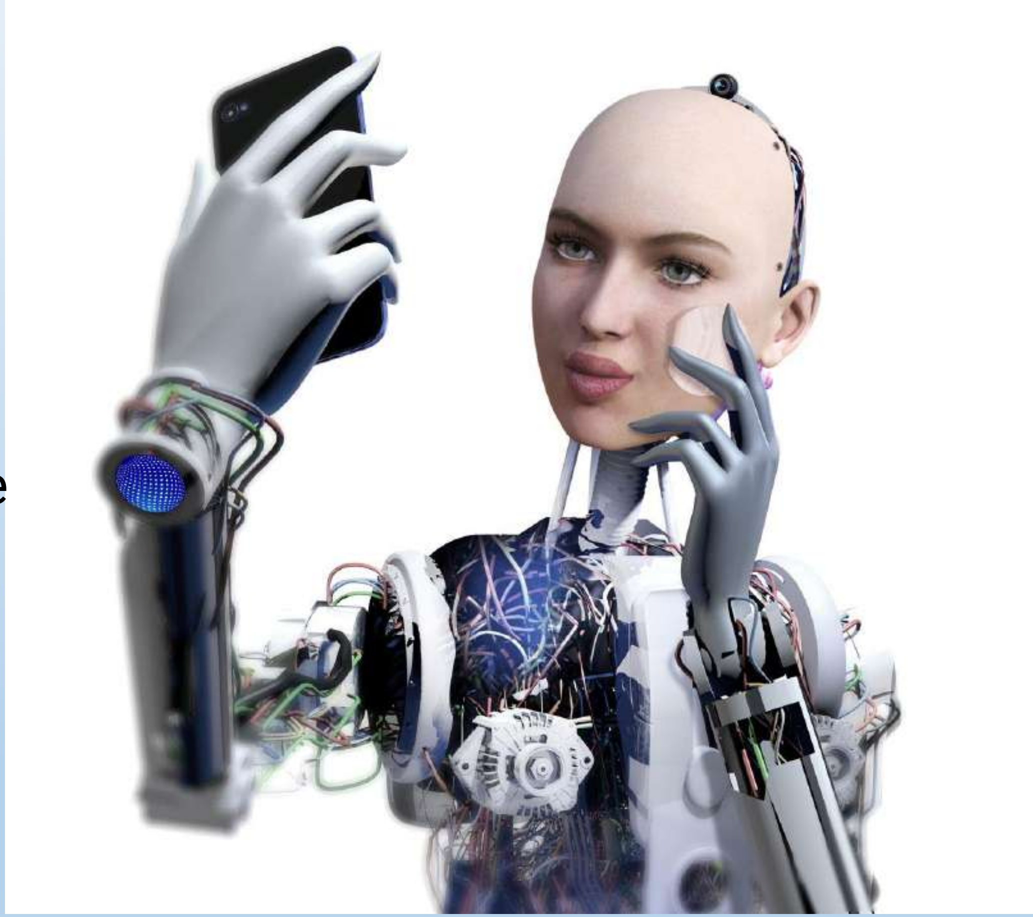
# Meet Microsoft twitter chatbot: Tay

*A chatbot is a form of AI which conducts a conversation via auditory or textual methods.*

released March 23 2016

learns from interacting with people on twitter

mimicks the language patterns of a 19-year-old American girl



shut down 16 hours after launch

official apology on Microsoft blog

Twitter ‘trolls’ took advantage of Tay's "repeat after me" capability by deliberately inputting offensive messages

*inflammatory and racist outputs from Tay*

Durham University

# COMPAS Algorithm: *Correctional Offender Management Profiling for Alternative Sanctions*

used in state court systems throughout the United States

predicts likeliness of criminal reoffending;

*Black defendants were almost twice as likely to be misclassified with a higher risk of reoffending (45%) in comparison to their white counterparts (23%).*

Durham University

# Facebook Ads

Ads tailored to demographic background

Facebook said that they have “made important changes”

*jobs such as nurses, secretaries and preschool teachers were suggested primarily to women*

*job ads for janitors and taxi drivers had been shown to a higher number of men, moreover men of minorities*

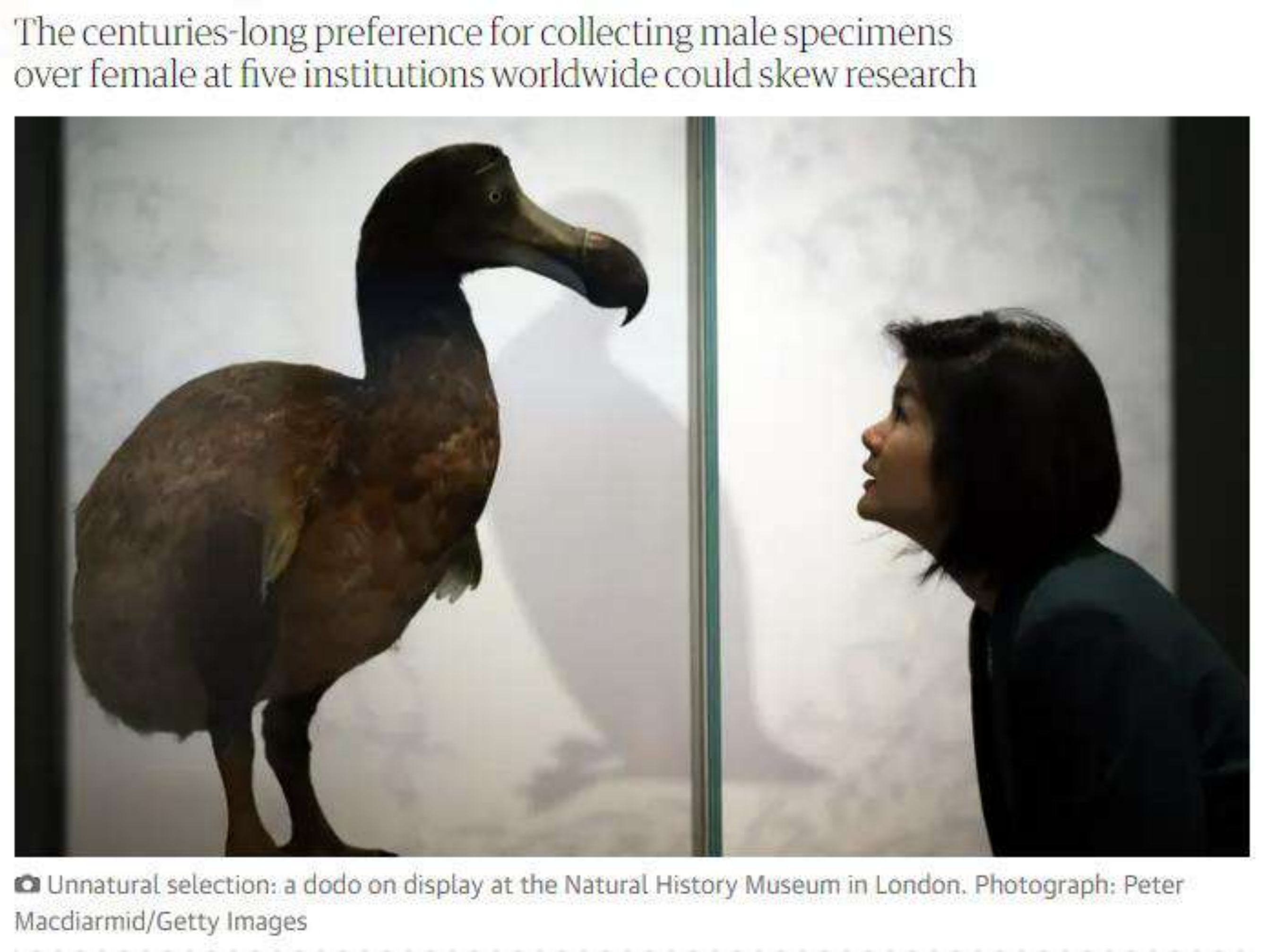
Durham University

# Bias in museums

- Why sexist bias in natural history museums really matters | Science | The Guardian …

Museums & Truth. The Truth is, there is More Than one Truth! - MuseumNext

a stereotypical museum culture which focuses on collecting and showcasing the stories, successes, and works of the white male in society.

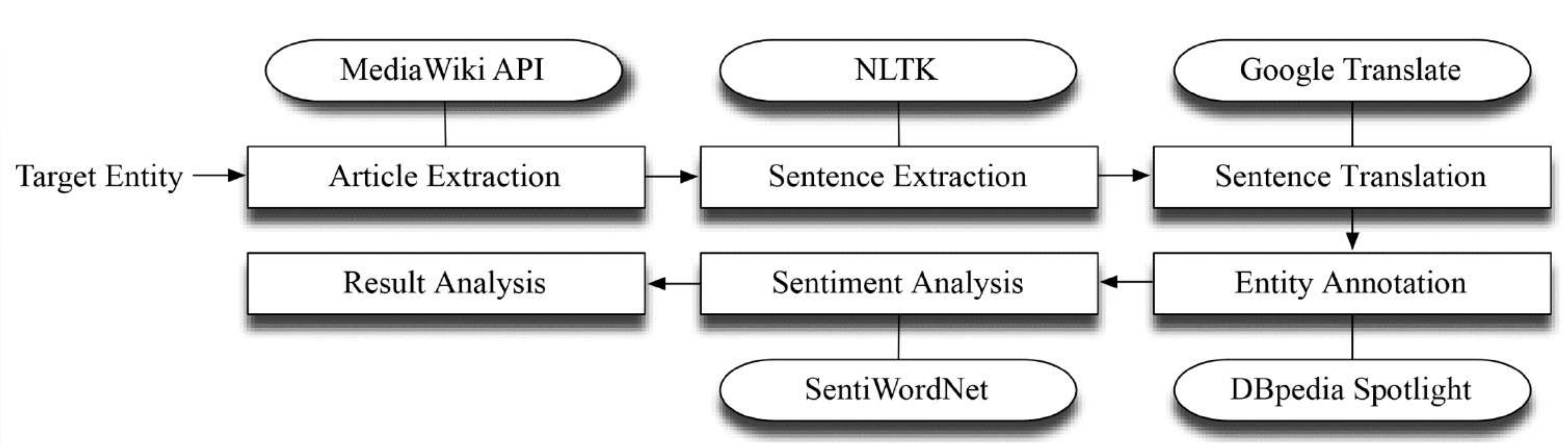


The centuries-long preference for collecting male specimens over female at five institutions worldwide could skew research

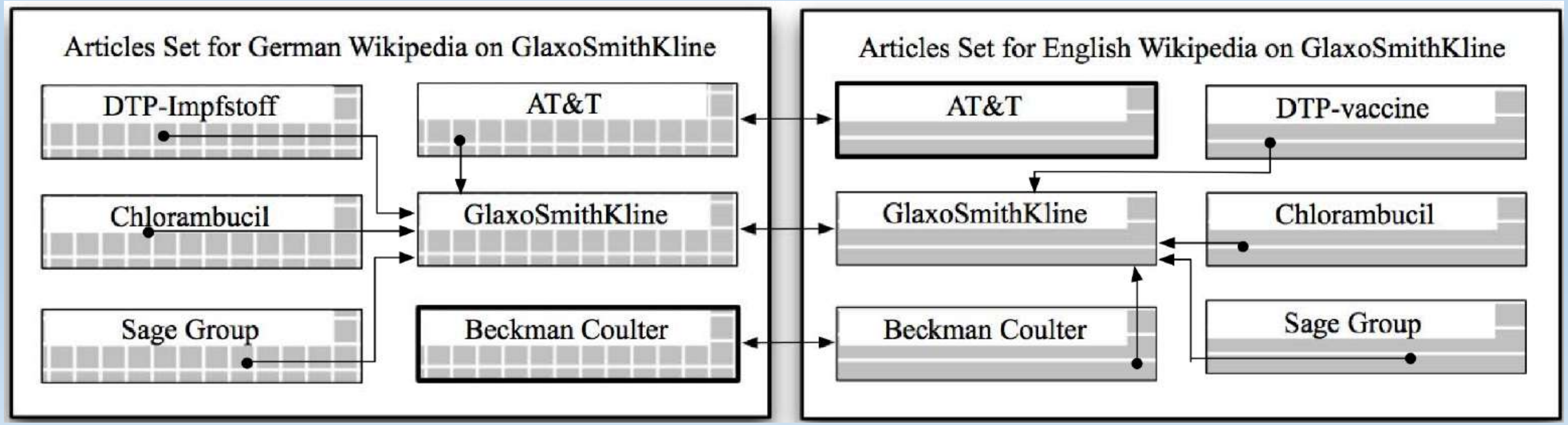
Unnatural selection: a dodo on display at the Natural History Museum in London. Photograph: Peter Macdiarmid/Getty Images

# Research Notes

# Wikipedia Study



- Zhou, Y., Demidova, E. and Cristea, A. I., 2016, “ Who likes me more? Analysing entity-centric language-specific bias in multilingual Wikipedia. ”, Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC 2016, Pisa, Italy, April 4-8, 2016).



# Wikipedia Study (NPOV)

- Angela Merkel Example:
  - majority of occurrences are located in German & English Wikipedia
  - success in the elections and some criticism she gets during her tenure
  - German Wiki:
    - Positives: compliments with respect to the time before she went on the political stage and became famous
    - Negatives: detailed personal information, regarding her haircut and clothes
  - English Wiki:
    - Relation to politicians, comments from other politicians about her
    - Positive: ‘I want to believe though, and I think I am right, that Angela Merkel is a fine leader with decent ethics and superior intelligence’
  - Portuguese Wiki:
    - Positive:  compliments to Angela Merkel’s performance in the economic crisis and on the financial market

Zhou, Y., Demidova, E. and Cristea, A. I., 2016, “ Who likes me more? Analysing entity-centric language-specific bias in multilingual Wikipedia. ”, Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC 2016, Pisa, Italy, April 4-8, 2016).

# Bias: Methodology & Evaluation

- Tsakalidis, A., Liakata, M., Damoulas, T., and Cristea, A. I., 2018, “ Can We Assess Mental Health through Social Media and Smart Devices? Addressing Bias in Methodology and Evaluation In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'18), Springer, 10-14 September 2018, Dublin, Ireland ( **Core A** )

**Training on past based on the future**.

**Overlapping instances across consecutive time windows**: biased if there are overlapping days of train/test data.

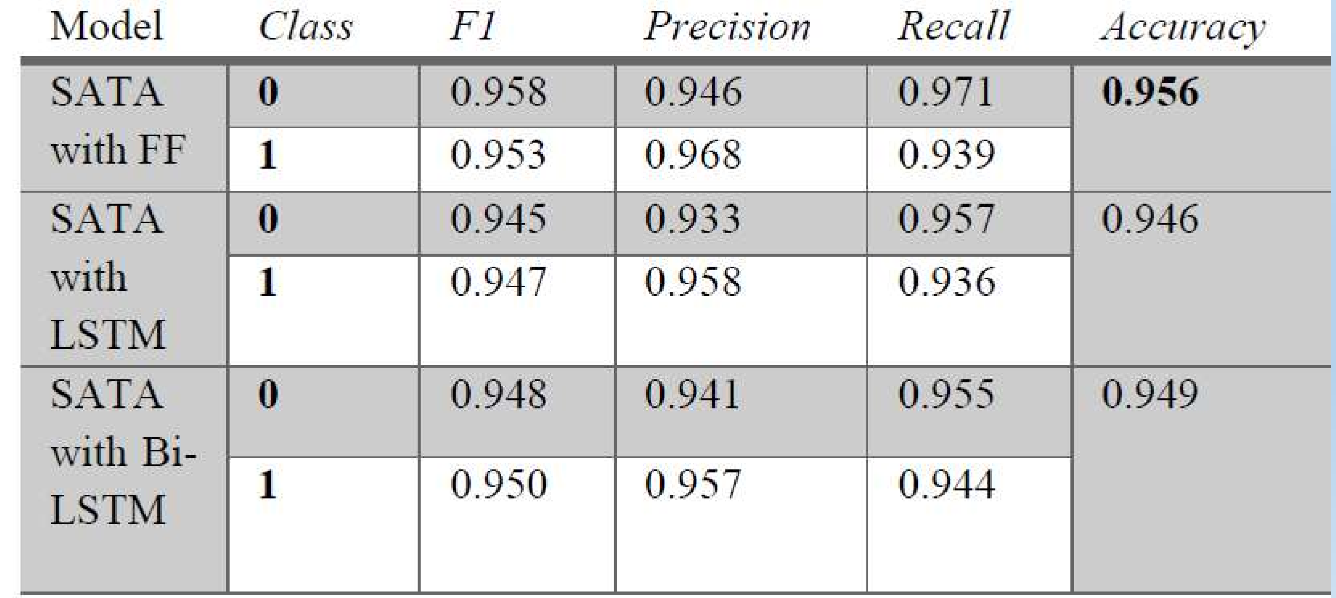
**Predicting users instead of mood scores**: most approaches merge all the instances from different subjects, in an attempt to build user-agnostic models in a randomised cross-validation framework

# Avoiding Bias

- Aljohani, Yu, J., T., Cristea, A. I., Author Profiling: Prediction of Learners' Gender on a MOOC Platform based on Learners' Comments, ICADMA 2020
- Bias in pre-course survey (due to unbalanced data or wrong inputs)
  - Solution: automatic profiling
- Bias in learning about the user instead of type of user
  - Solution: different users in training and test sets
- Bias in future data predicting past
  - Solution: training on past, testing on future
- Bias in unbalanced data sample
  - Solution: stratified sampling (homogenous groups by label); text augmentation (paraphrasing)

# Avoiding Bias

- Aljohani, Yu, J., T., Cristea, A. I., Author Profiling: Prediction of Learners’ Gender on a MOOC Platform based on Learners’ Comments, ICADMA 2020
- Bias in Methodology – average accuracy
  - Solution: results given per class



| Model | *Class* | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|---|
| SATA with FF | **0** | 0.958 | 0.946 | 0.971 | **0.956** |
| | **1** | 0.953 | 0.968 | 0.939 | |
| SATA with LSTM | **0** | 0.945 | 0.933 | 0.957 | 0.946 |
| | **1** | 0.947 | 0.958 | 0.936 | |
| SATA with Bi-LSTM | **0** | 0.948 | 0.941 | 0.955 | 0.949 |
| | **1** | 0.950 | 0.957 | 0.944 | |

# Further Papers

- J. Yu *et al*., "INTERACTION: A Generative XAI Framework for Natural Language Inference Explanations," *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892336.

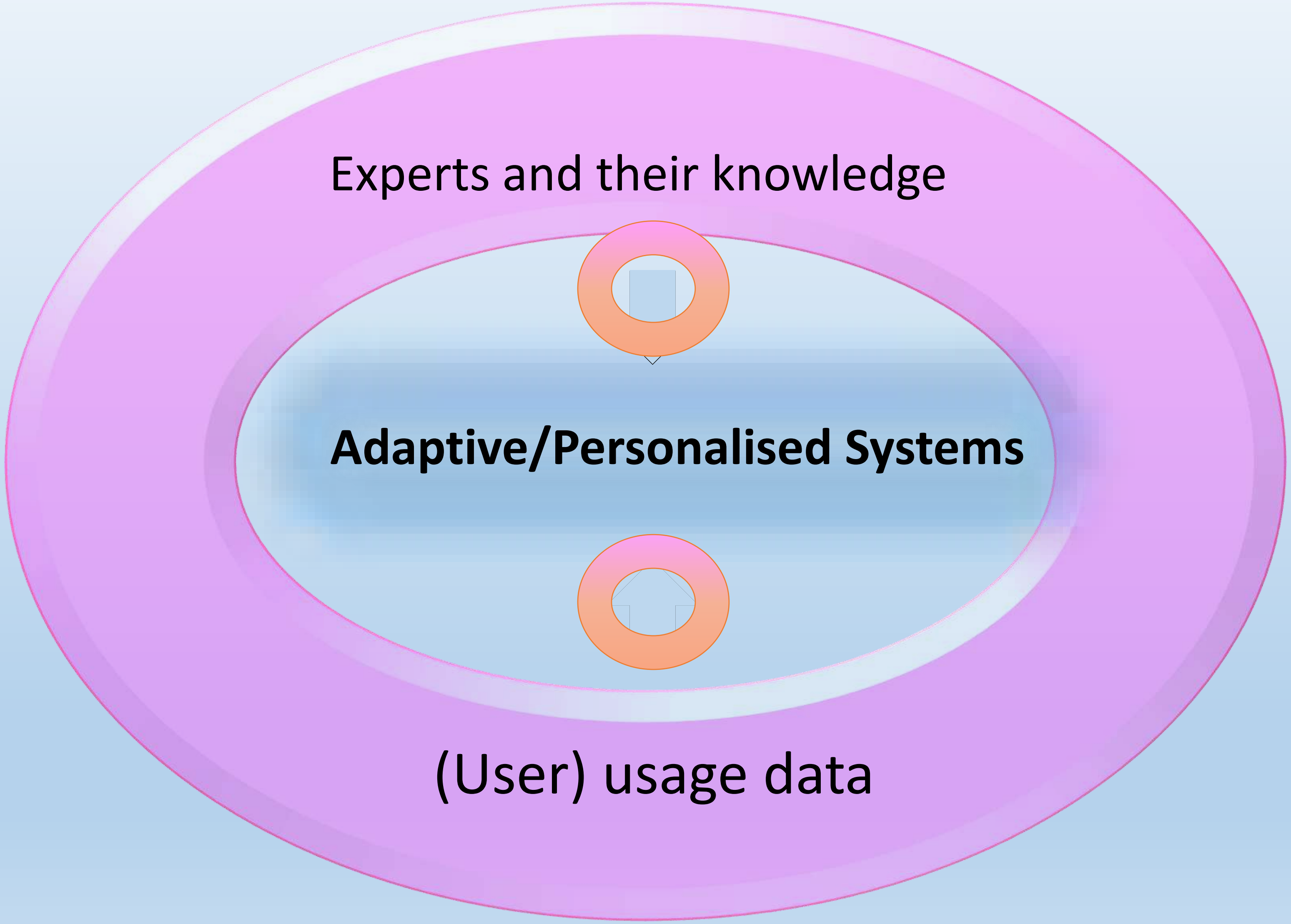
# Categories of bias in AI

- Researchers have identified three categories of bias in AI:
- *Algorithmic prejudice* occurs when there is a statistical dependence between protected features and other information used to make a decision.
- *Negative legacy* refers to bias already present in the data used to train the AI model.
- *Underestimation* occurs when there is not enough data for the model to make confident conclusions for some segments of the population.

Durham University

# Solutions & The Future

# AI: Top Down versus Bottom Up



Experts and their knowledge

**Adaptive/Personalised Systems**

(User) usage data

Durham University

# Conclusions & food for thought

- Unique opportunity?
- Future endangered?

Alexandra I Cristea - Google Scholar

Durham University