Durham University

Keynote
Professor Alexandra I. Cristea

Professor of Computer Science

Bias in AI

Artificial Intelligence is a thriving area in Computer Science. Especially trending is the sub-area of Machine Learning and Deep Learning, including Data Analytics. However, the latter comes often with various forms of bias. Bias in AI can be introduced in many forms, from data to methods and algorithms, and it negatively affects people as well as research quality. It also impacts upon an increasing amount of areas, including sensitive ones, such as healthcare, law, criminal justice, hiring. Thus, an important task for researchers is to use AI to identify and reduce (human or machine) biases, as well as improve AI systems, to prevent introducing and perpetuating bias.

Durham University

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

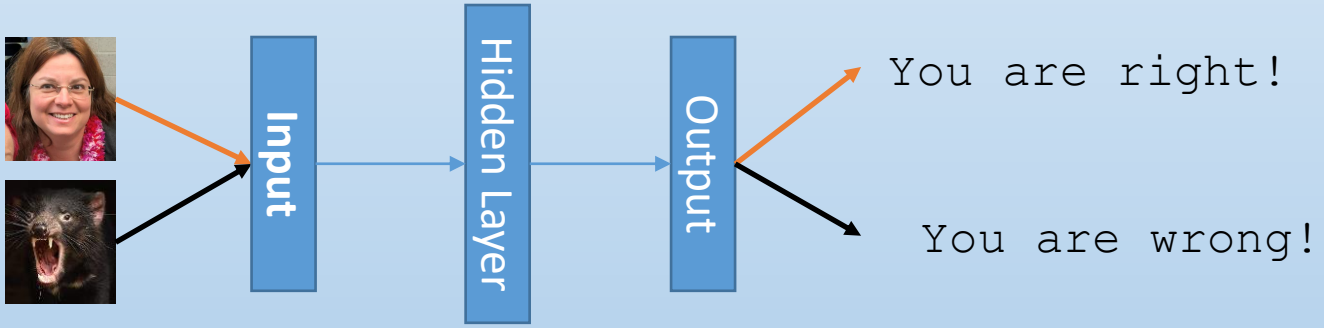
- 'black-box' deep NN:

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



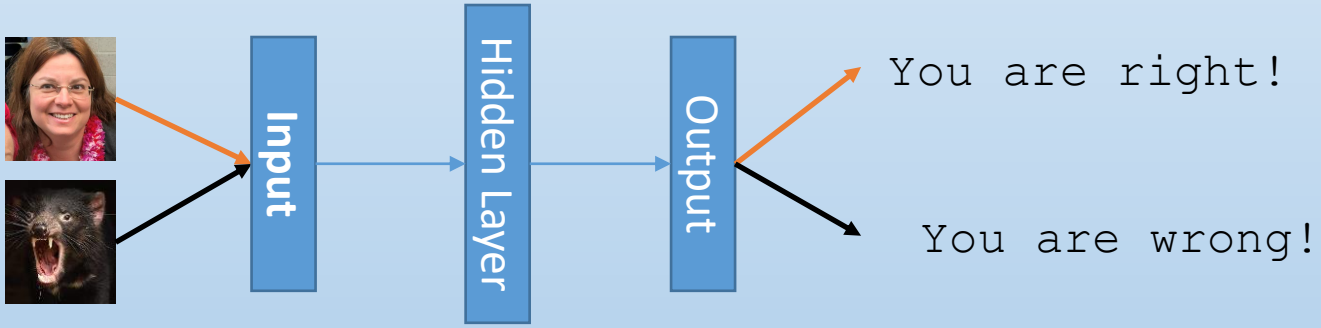
- 'black-box' deep NN:

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



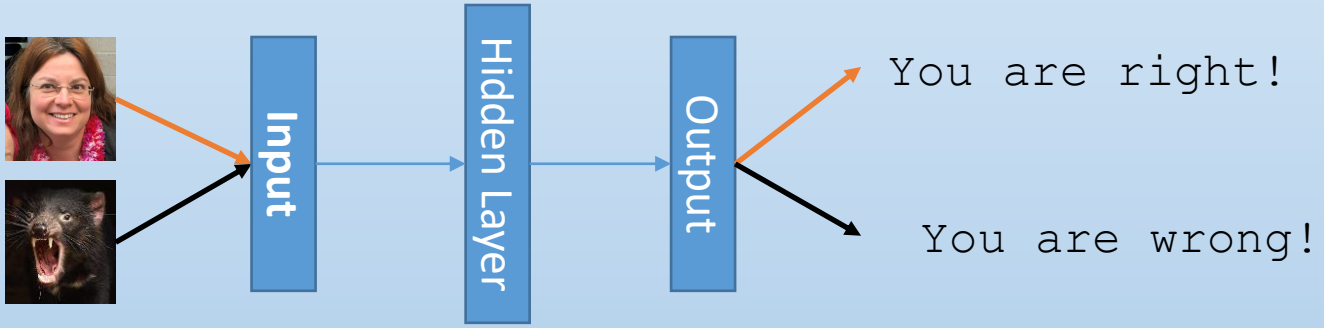
- 'black-box' deep NN: train on

What is bias in AI?

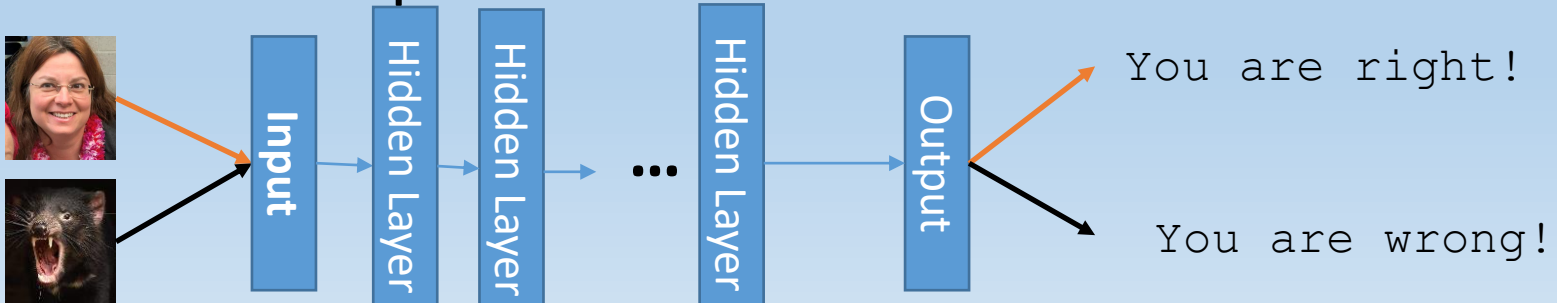
- Explicit, rule-based AI:

```
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: train on



Durham University

Bias in real life

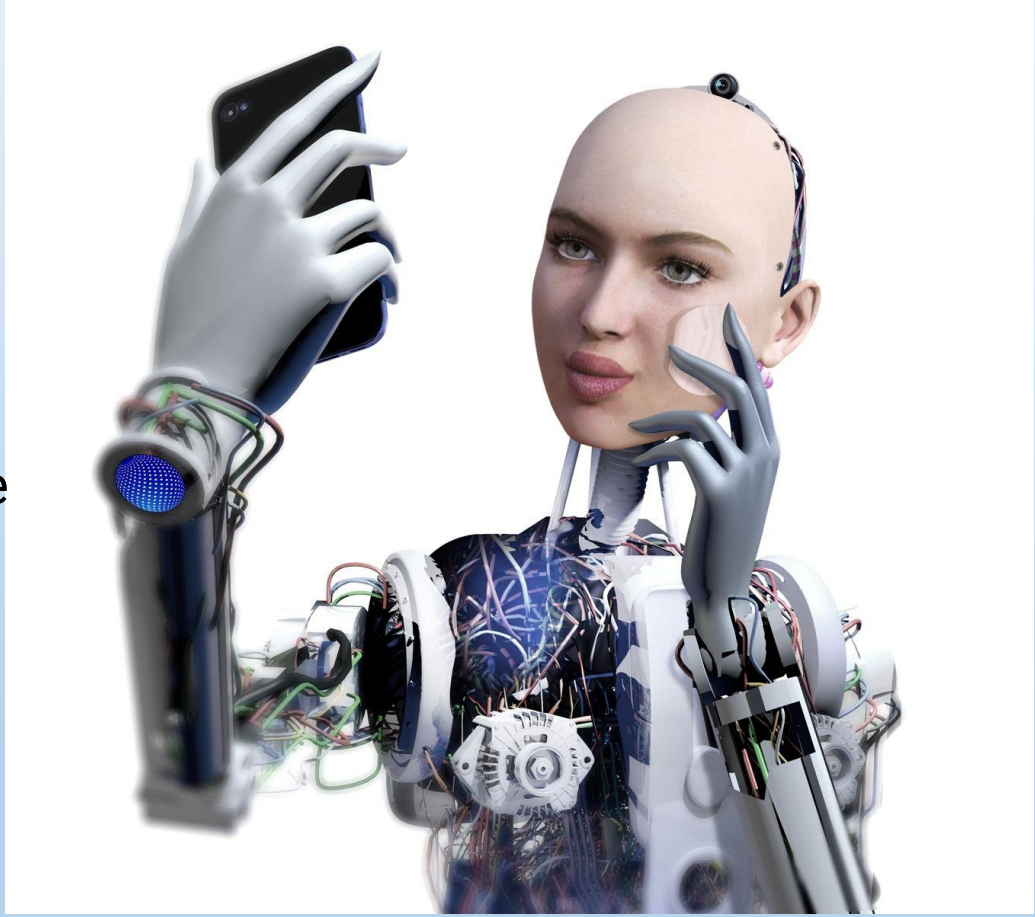
Meet Microsoft twitter chatbot: Tay

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

GIPHY: Gender classification via iris information

machine learning algorithms can work out someone's gender from a picture of their iris

images of eyes with and without eyeliner

gender from eye makeup?

Durham University

Facebook translation

- in October 2017 the Israel police mistakenly arrested a Palestinian after relying on automatic translation software. The service translated a picture of the construction site worker "good morning" as "attack them".

Durham University

Apple’s new credit card

Apple’s new credit card may give higher limits to men than to women

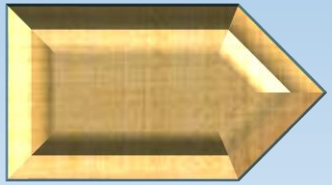
Goldman Sachs, which issues the card, said its credit decisions were “based on a customer’s creditworthiness and not on factors like gender, race, age, sexual orientation or any other basis prohibited by law.”

Durham University

Google & Amazon

- artificial intelligence services from Google and Amazon both failed to recognize the word “hers” as a pronoun, but correctly identified “his.” (Nov 11, R. Munro)

Today, “hers” is not recognized as a pronoun by the most widely used technologies for Natural Language Processing (NLP), including (alphabetically) Amazon Comprehend, Google Natural Language API, and the Stanford Parser.



Bias in archives, libraries

- List of statements on bias in library and archives description – Cataloging Lab
- Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). [Sensitivity message appears as a pop-up with information about language used in resources]
- Australian War Memorial. Disclaimer [along with pop-up with information about language used in resources]
- Brown University Library. Terminology [statement on African American history description]
- …

Silence and Bias in Archives - Archives and Special Collections - Research Guides at Ryerson University Library

Archival silences refer to the erasure of archives, and histories of marginalized communities within traditional archival holdings.

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

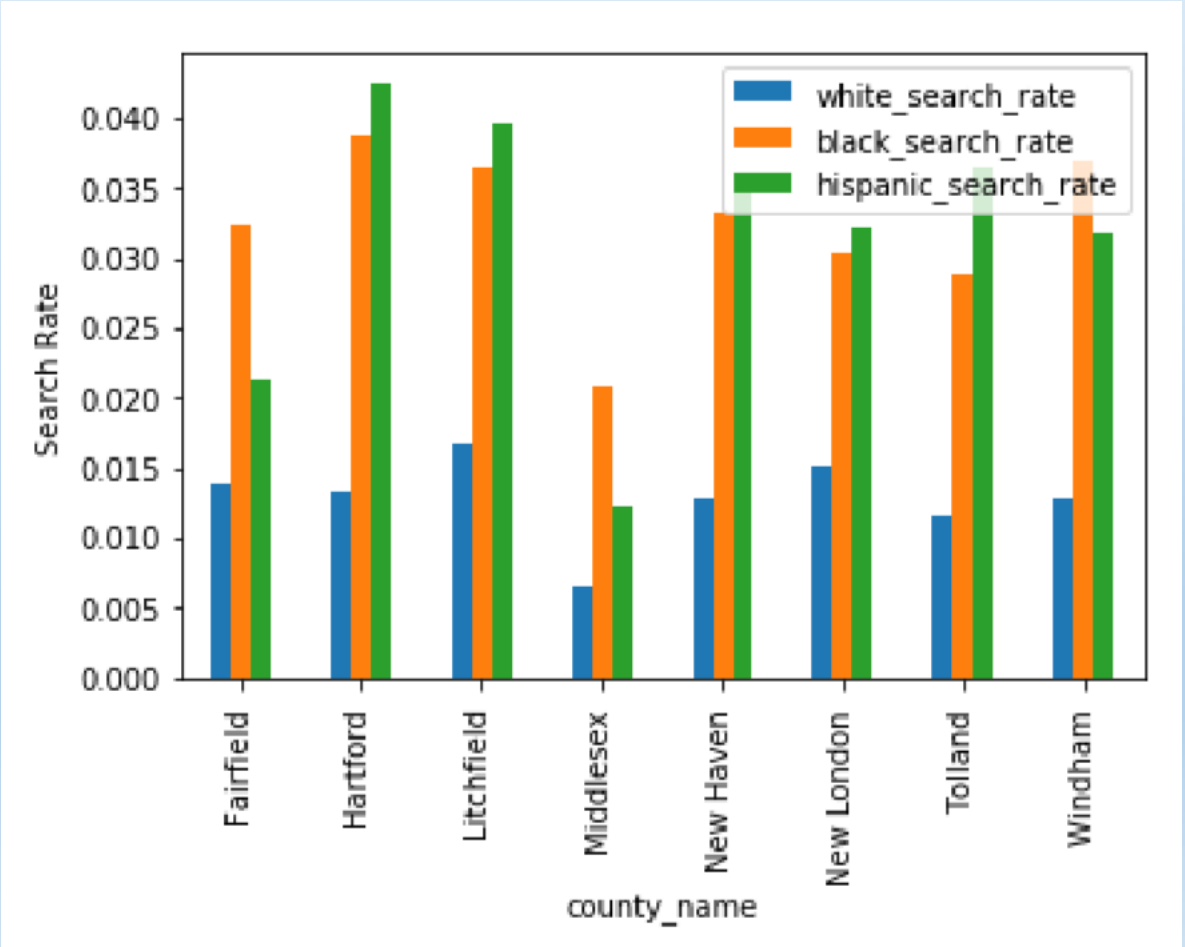
Unnatural selection: a dodo on display at the Natural History Museum in London. Photograph: Peter Macdiarmid/Getty Images

Durham University

Research Notes

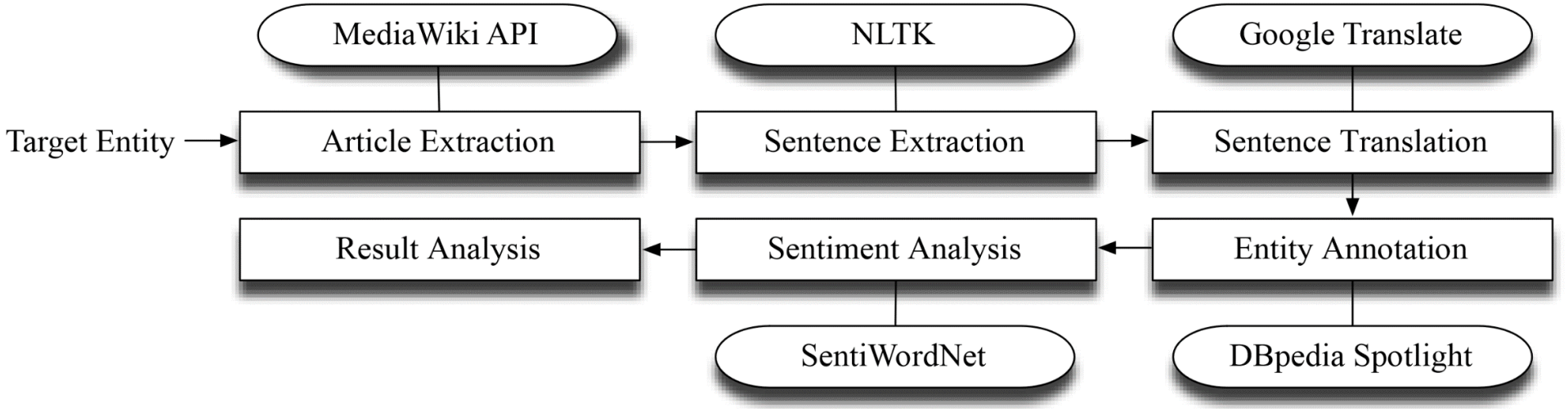
Bias Research in AI is not new

- James J. Heckman, Sample Selection Bias as a Specification Error, Econometrical, 47(1), Jan 1979 (cited by 29844 in Google Scholar)

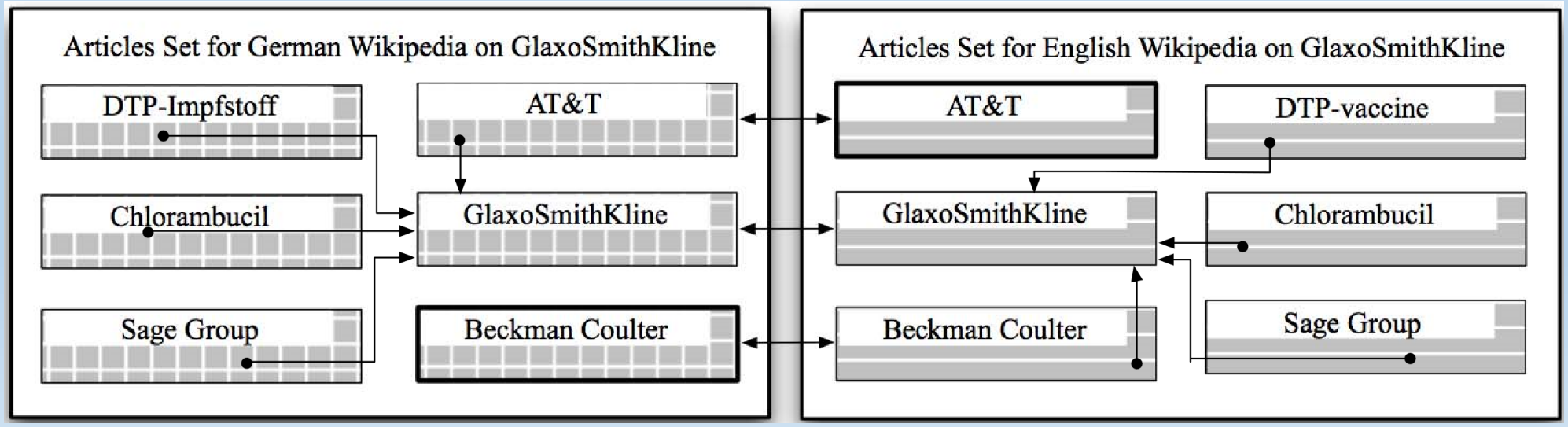


Durham University

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).



Durham University

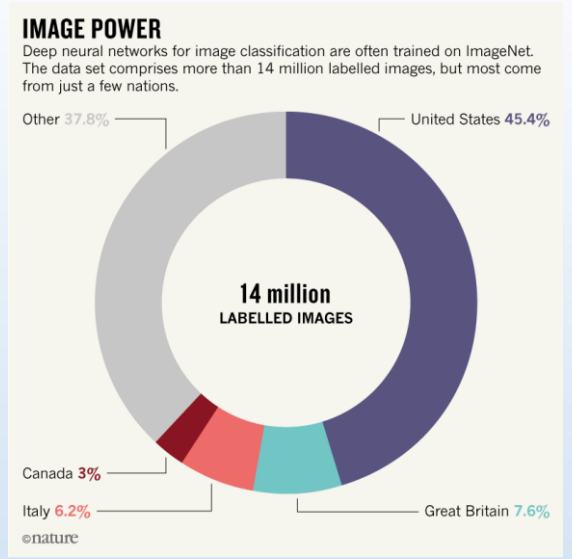
Wikipedia Study (NPOV)

- 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).

Durham University

Bias: Skewed input data

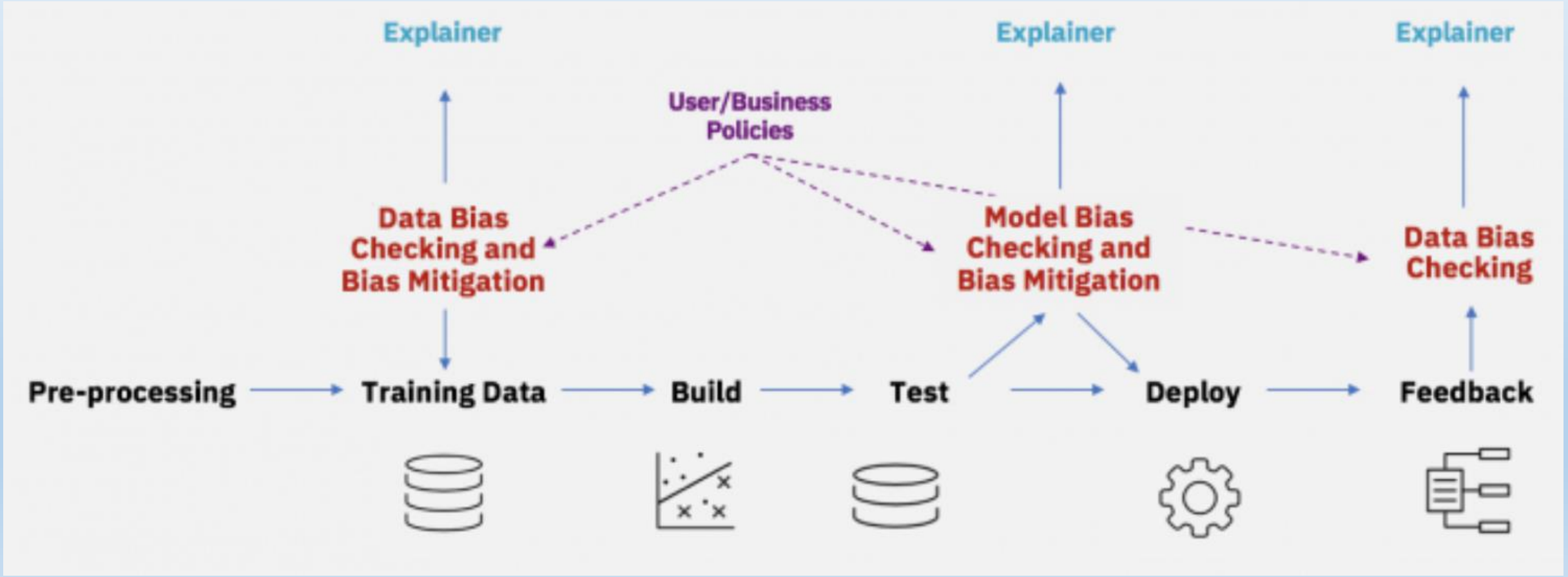


- Nature, 2018: https://www.nature.com/articles/d41586-018-05707-8
- ML trained on large, annotated data sets (ImageNet, a set of more than 14 million labelled images; NLP: corpora of billions of words)
- Sources: Google Images, Google News, w. specific query terms; Wikipedia. Annotated via e.g. Amazon Mechanical Turk.
- Issue:
  - some groups over-represented, others are under-represented.
  - > 45% of ImageNet data, is from US, (4% world population). China & India contribute 3% of ImageNet data & represent 36% of the world's population.

Durham University

Bias in Queries e.g. percentages

- AI FAIRNESS 360 by IBM Research, 2018



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

Durham University

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 (to appear)
- 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	

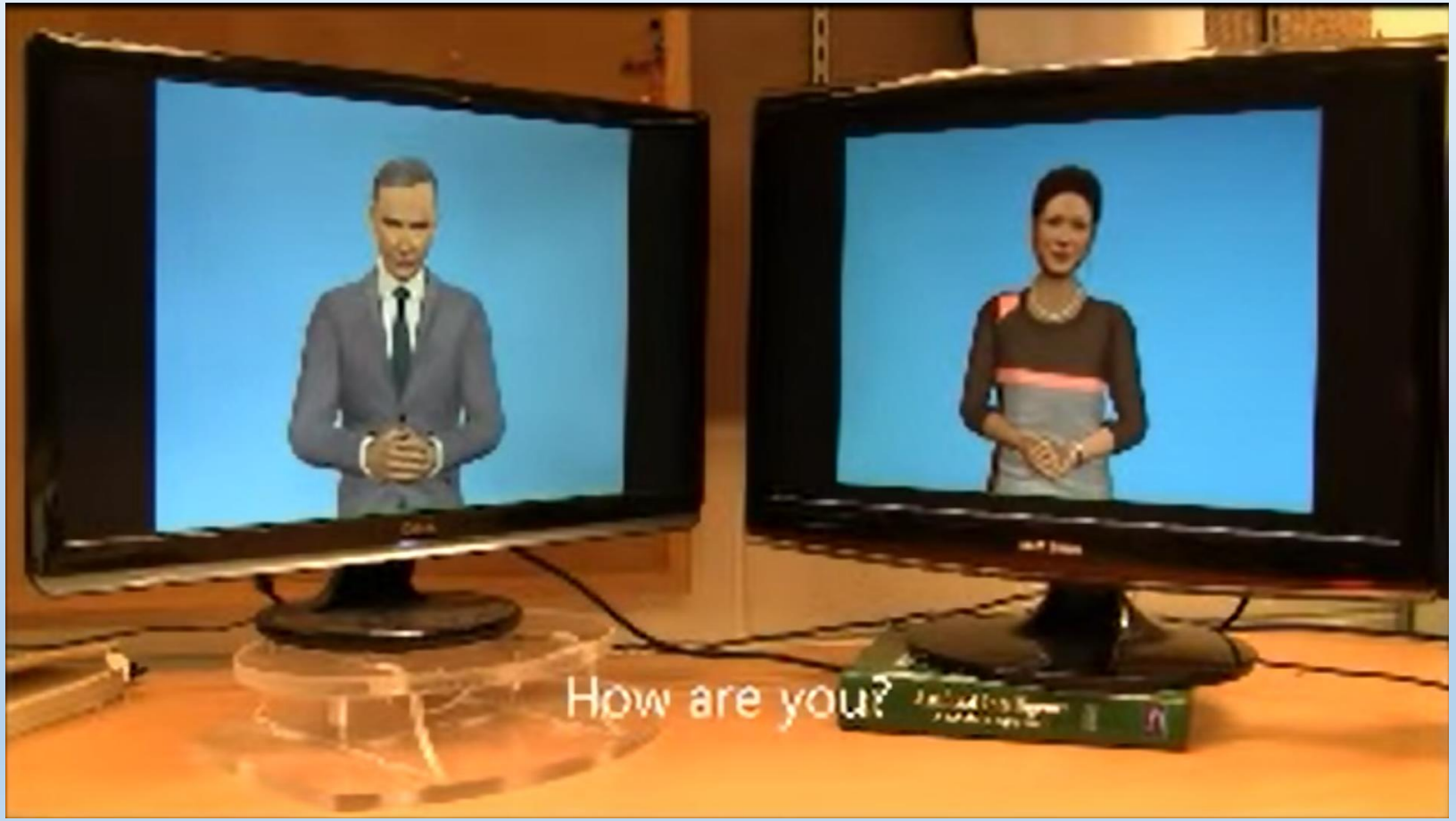
Categories of bias in AI

- 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.

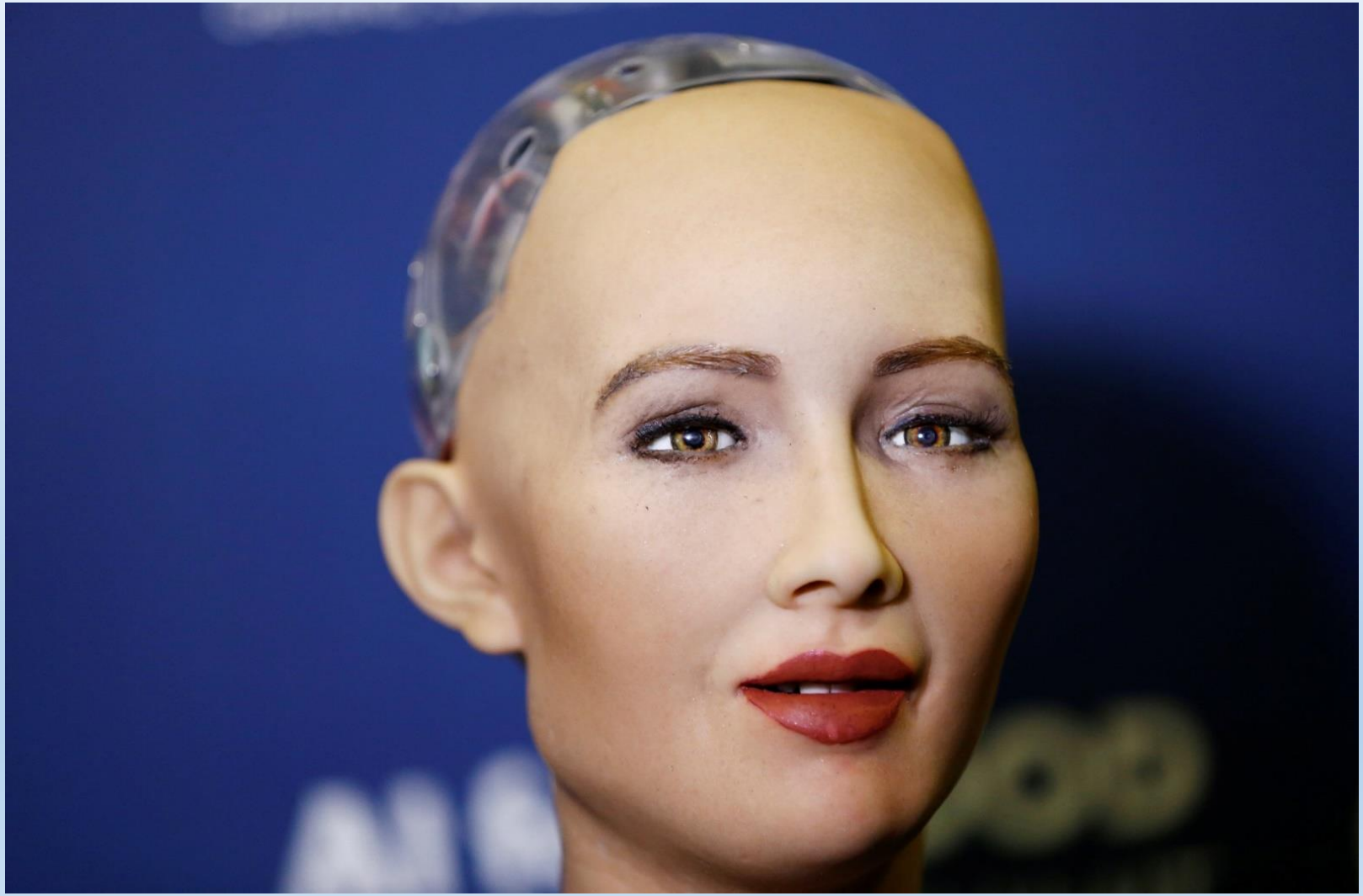
Durham University

AI & Society
Social Impact

AI versus AI: Chatbots quarrelling



Meet Sophia, first AI citizen of a country



Durham University

Durham University

Solutions & The Future

IBM Research:

- ***AI bias will explode. But only the unbiased AI will survive.***

- *Within five years, the number of biased AI systems and algorithms will increase. But we will deal with them accordingly – coming up with new solutions to control bias in AI and champion AI systems free of it.*

Durham University

New York Times

"All the News That's Fit to Print"

The New York Times

LATE CITY EDITION

VOL. CXVIII . No. 40,721 NEW YORK, MONDAY, JULY 21, 1969 10 CENTS

MEN WALK ON MOON

ASTRONAUTS LAND ON PLAIN; COLLECT ROCKS, PLANT FLAG

Voice From Moon: 'Eagle Has Landed'

A Powdery Surface Is Closely Explored

By JOHN NOBLE WILFORD

- ***The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.*** (The New York Times, Nov 15th 2019)

Durham University

Actions (as per Harvard Business Review)

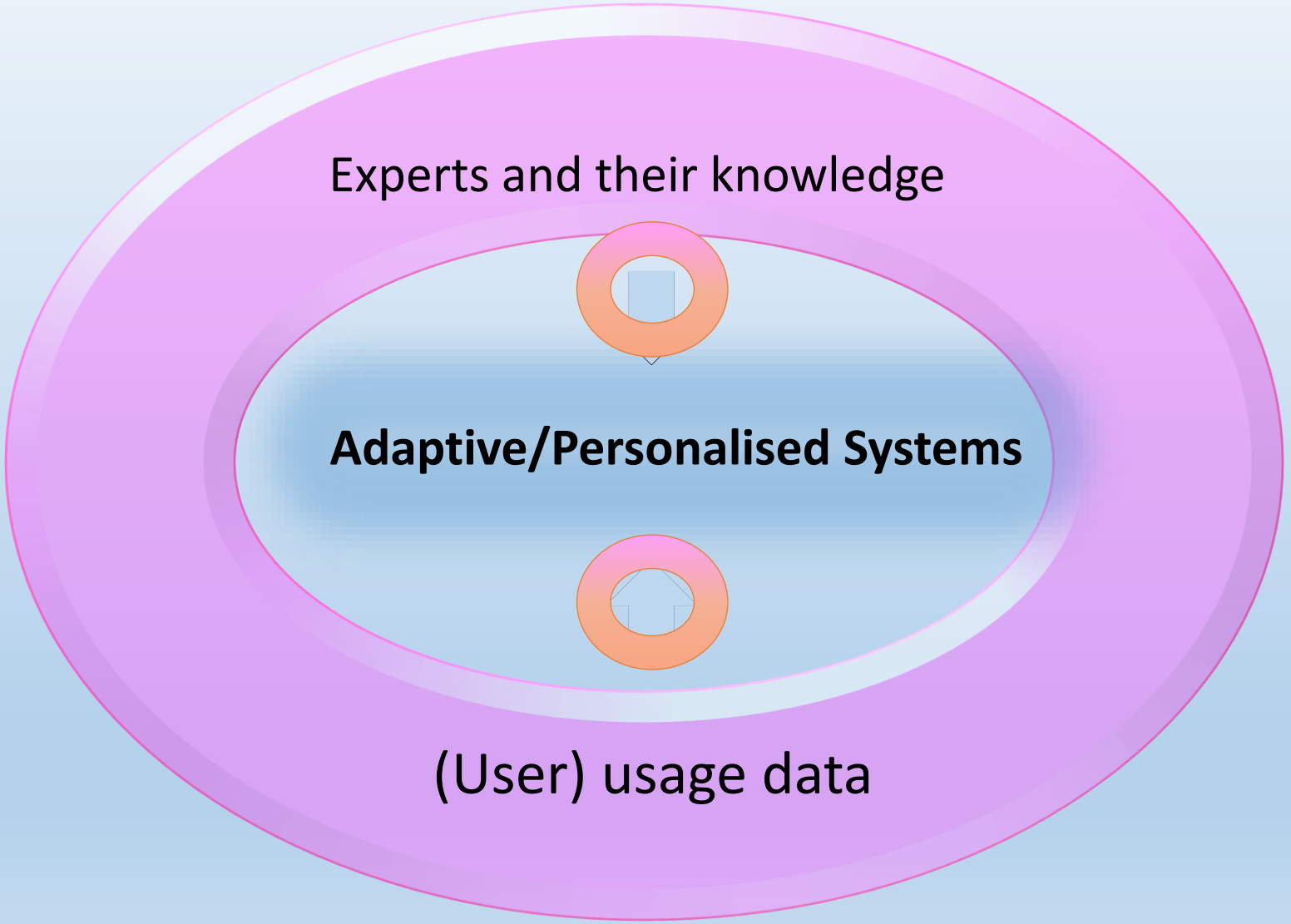
- Take responsibly advantage of the way

  AI can improve on human decision-making

  - Unconscious bias, lies, human brain as ‘black box’
- Accelerate progress in addressing Bias in AI
  - Google AI has published recommended practices
  - IBM ‘Fairness 360’ framework with technical tools
  - Explainable techniques

Durham University

AI: Top Down versus Bottom Up



Durham University

Addressing bias in AI (for the practician)

- Define and narrow the business problem you're solving.
- Structure data gathering that allows for different opinions.
- Understand your training data.
- Gather a diverse ML team that asks diverse questions.
- Think about all of your end-users.
- Annotate with diversity.

Conclusions & food for thought

- Unique opportunity?
- Future endangered?

Durham University