Housekeeping

- Please mute your mic during the talk
- There will be time at the end for discussion in chat, publically or privately
- This session will be recorded
- Slides and recording will be available after
- Follow on twitter @equal4success

Babraham Institute

Biases in AI

Izzy Hampson

Babraham Institute

Definition

- **AI** (artificial intelligence) is the leading technology in “Fourth Industrial Revolution”. It refers to the technological advances – from biotechnology to big data – that are rapidly reshaping the world as we know it.

Babraham Institute

Artificial Intelligence has a gender bias problem… just ask Alexa

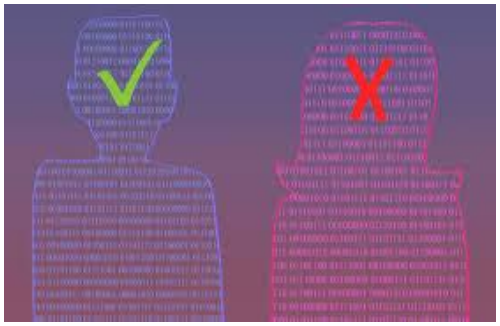
- Suggest to Samsung's Virtual Personal Assistant Bixby "Let's talk dirty".

- "I don't want to end up on Santa's naughty list."
- "I've read that soil erosion is a real dirt problem."

Babraham Institute

Bias and discrimination in AI

- AI systems are often biased, particularly along race and gender lines.
- Amazon recruitment algorithm displayed gender biases.
- The algorithm was trained on historical data and the preferential recruitment of males, it ultimately could not be fixed and had to be dropped.



amazon

Babraham Institute

Gender Biases

- Link between development of AI systems with gender biases and the lack of women in design teams.
- **AI Now** report: clear connection between the male dominated AI industry and its discriminatory systems and products.
- Less recognition of the ways AI products incorporate stereotyped **representations** of gender in their design.

AINOW INSTITUTE

Babraham Institute

Word-embeddings

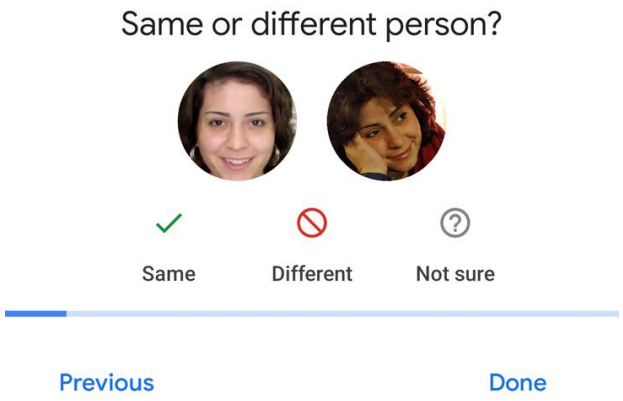
- AI fills in the word **queen** in the sentence *"Man is to king, as woman is to X."*
- Words are converted to numerical representations used as inputs models.
- Words are represented as a sequence, or a vector of numbers. If two words have similar meanings, their "embeddings" will be mathematically close to each other.
- The issues arise in cases where AI fills in sentences like *"Father is to doctor as mother is to nurse."*

Project Implicit®

Babraham Institute

Racial Bias in AI

- Chatbots that become racist in less than a day
- Facial technology that fails to recognize users with darker skin colours
- Ad-serving algorithms that discriminate by gender and race
- An AI hate speech detector that's racially biased itself



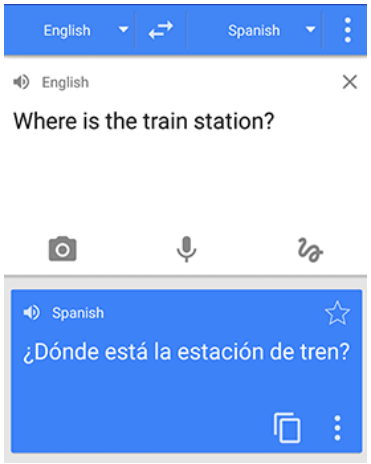
Babraham Institute

- Racial biases are **thoroughly ingrained** in society, and have the potential to be exacerbated by algorithms, such as in the criminal justice system.
- Significant problems include the lack of unbiased historical data, an unbalanced workforce, and limited user testing.

Babraham
Institute

What causes AI bias?

- An incomplete or skewed training dataset
- Labels used for training
- Features and modelling techniques

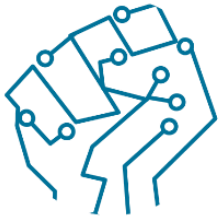


The Bride Problem:

Babraham Institute

Ways to address bias in AI

- Ensure diversity in the **training samples**
- Ensure that humans **labelling** the samples come from diverse backgrounds.



Black in AI

The Alan Turing Institute

- Encourage machine-learning teams to measure accuracy levels separately for different demographic categories and to **identify** when one category is being treated unfavourably e.g. Counterfactual Fairness
- Collect **more training data** associated with underrepresented groups and apply machine learning de-biasing techniques that can penalise for producing unfairness.

Babraham Institute

*"Bias is all of our **responsibility**. It hurts those discriminated against, of course, and it also hurts everyone by reducing people's ability to participate in the economy and society. It reduces the potential of AI for business and society by encouraging mistrust and producing distorted results. Business and organizational leaders need to ensure that the AI systems they use improve on human decision-making, and they have a responsibility to encourage progress on research and standards that will reduce bias in AI."*

-James Manyika, Jake Silberg and Brittany Presten, 2019, Harvard Business Review

Babraham Institute

Resources

- Man is to Computer Programmer as Woman is toHomemaker? Debiasing Word Embeddings (https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf)
- Gender Shades: Intersectional Accuracy Disparities inCommercial Gender Classification (http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf)
- AI and Gender Four Proposals for Future Research (http://lcfi.ac.uk/media/uploads/files/AI_and_Gender_4_Proposals_for_Future_Research.pdf)
- DISCRIMINATING SYSTEMS: Gender, Race, and Power in AI (https://ainowinstitute.org/discriminatingsystems.pdf)
- Harvard implicit association test (https://implicit.harvard.edu/implicit/)

Babraham Institute