# SICCS: Algorithmic Bias, Fairness, & Justice

**Taught Seminar: May 30 2025**

Zee Talat

ztalat@ed.ac.uk

THE UNIVERSITY of EDINBURGH Edinburgh Futures Institute

THE UNIVERSITY of EDINBURGH informatics

# Part I
# **Fairness and Ethics**

# What does it mean to be ethical?

- ~2500-3000 years of people asking the question: How do we lead a good life?

# What does it mean to be ethical?

- ~2500-3000 years of people asking the question: How do we lead a good life?
- To be precise and accurate
- To be accountable to the methods and outcomes of your work

# Why care about ethics?

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia

Nazi human experimentation

# Why care about ethics?

# Why care about ethics?

- Physical impacts are a subset of all negative impacts of work
- CSS is particularly prone to dual use concerns
  - By dual use we mean
    *"the malicious reuse of technical artefacts developed without harmful intent. Malicious reuse denotes applications that are used to harm any, and in particular marginalized groups in society."*
- CSS is particularly prone to concerns about surveillance

# How do we do ethical work?

- Ensure methods match research questions/purposes
- Make sure that your questions are meaningful and valid
- Make sure the outcomes of your work are not harmful
- And that releasing your work/using it does not cause harm

# What are ways CSS methods can harm?

- Two types of harms:
  - Representational harms
  - Allocative harms

# What are ways CSS methods can harm?

- Language technologies are always socially biased
- Network methods can cause harms, in particular, through surveillance
- Data Science can cause harm by correlating actors with actions/attributes

# Question Time

# So What Can I Do?

- Prevent others from engaging unethically with your work
- Not engage unethically with your work

# But I can’t control other people?

- Limit the release of your data/code to verified actors
- Specify how the data/methods can be used
- And what purposes they can and cannot be used for

# Okay but what about us?

- Bias measurement
- Actionability
- Acountability
- Transparency
- Measurement Validity
  - Construct Validity
- Interpretability
- Explainability

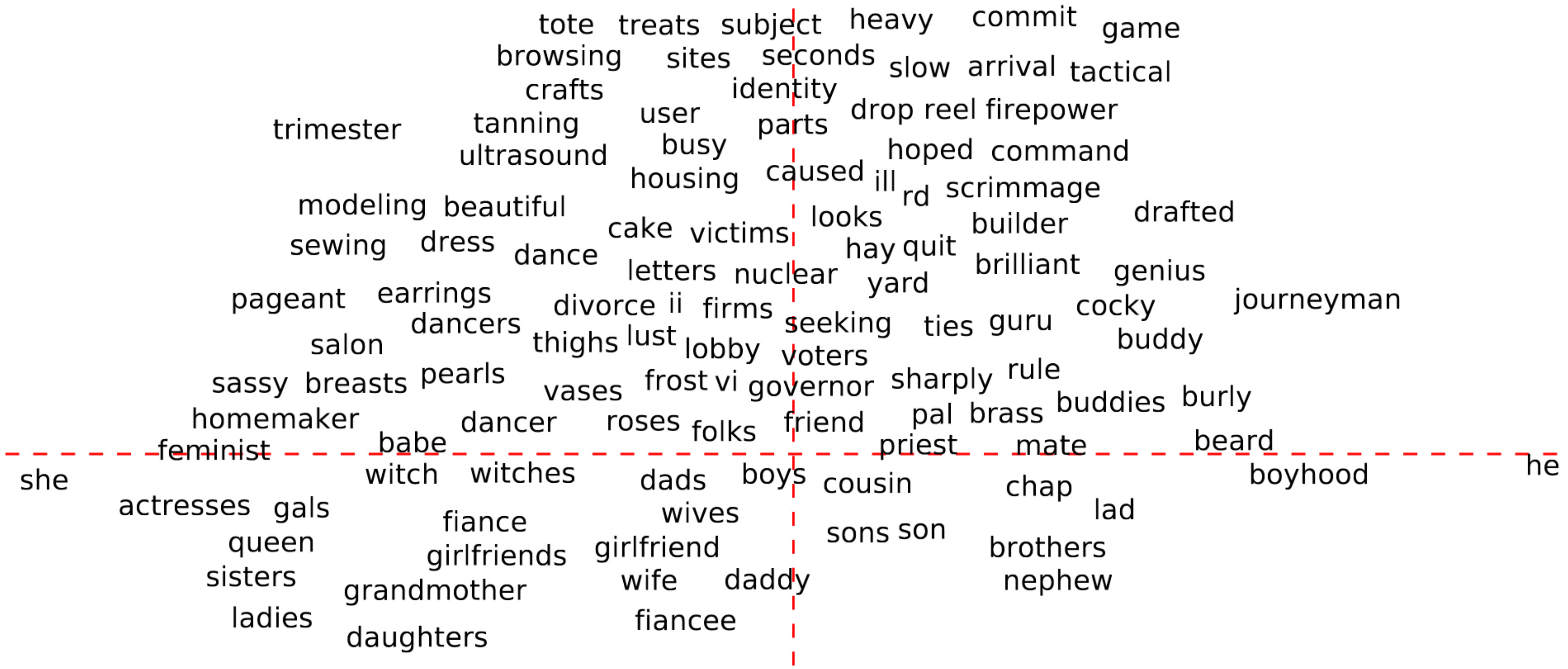
# Methods for Fairness

- Extrinsic evaluations
  - E.g., loan applications
  - Tying demographic background closely into the work
- Intrinsic Evaluations
  - E.g., what the internal model representations look like

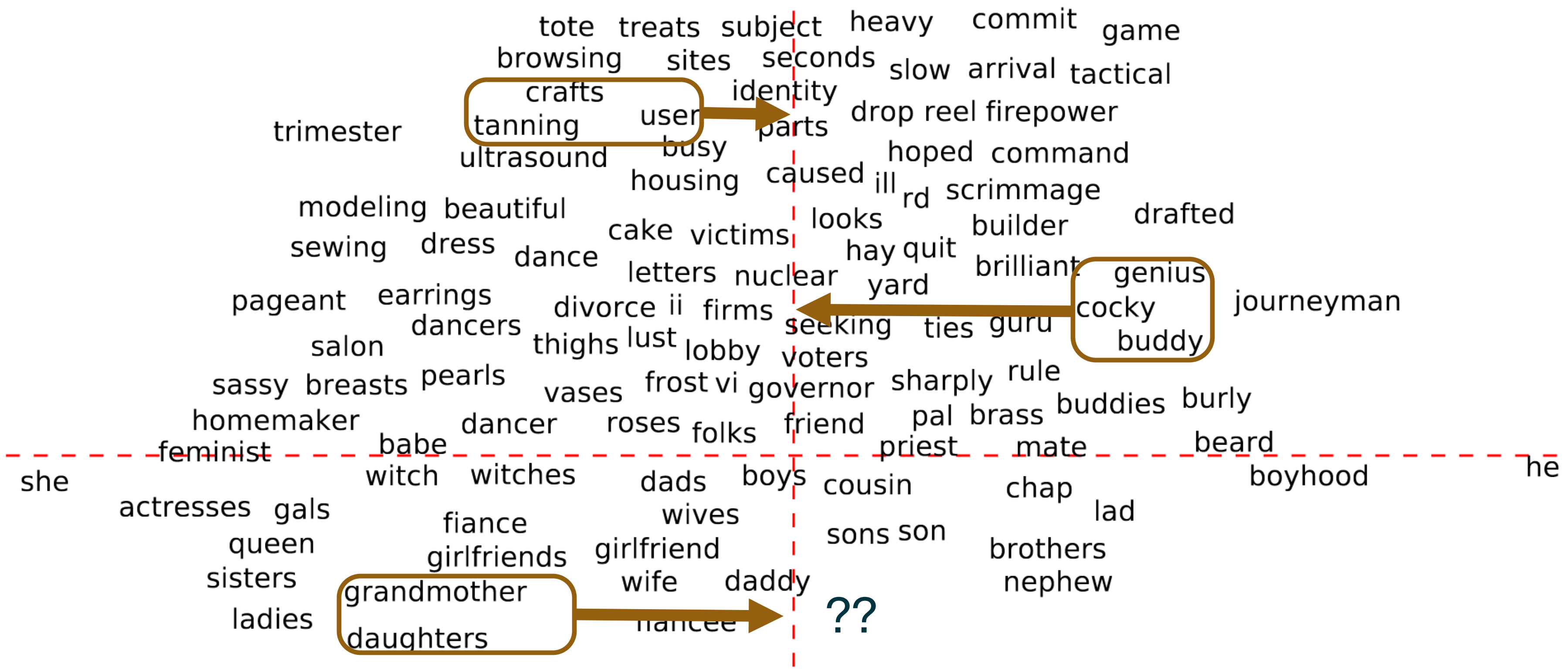
# Documenting Poor Performances

- Bad systems can be used to enact harms on people
  - E.g., misrecognition from facial recognition technology
- But are good systems good?
  - Is a good facial recognition system a good thing?
  - Do we want to help improve such systems to perform better?

# Methods in NLP



# Methods in NLP




THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

THE UNIVERSITY of EDINBURGH
informatics

# It does... not work

- Level 1 of not working
- Level 2 of not working

# Is that all we can do?

- Fine-tuning for good
- “Participatory” approaches
  - Red teaming
  - Reinforcement Learning with Human Feedback

# Question Time

# Usability of Bias Evaluation Metrics

*"Actionability refers to the degree to which a [bisa] measure's results enable decision-making or intervention; that is, results from actionable bias measures should facilitate informed actions with respect to the bias under measurement." – Delebolle et al. (2024)*

# Usability of Bias Evaluation Metrics

*"Actionability refers to the degree to which a [bisa] measure's results enable decision-making or intervention; that is, results from actionable bias measures should facilitate informed actions with respect to the bias under measurement." – Delebolle et al. (2024)*

# Desiderata for Actionability

***We want clarity(!) of***

- Motivation for the bias measure
- The underlying bias construct
- Intervals and ideal results
- Intended uses
- Reliability

# Accountability

- Accountability is for “establish[ing] informed and consequential judgments of… AI systems”
  - *Birhane et al., 2024. “AI auditing: The Broken Bus on the Road to AI Accountability.”*
- And for ensuring that “responsible or answerable for a system, its behavior and its potential impacts”
  - *Raji et al., 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing.*
- However, “AI audit studies do not consistently translate into more concrete objectives to regulate system outcomes.”
  - *Birhane et al., 2024. “AI auditing: The Broken Bus on the Road to AI Accountability.”*

# Transparency

- Transparency is about “what information about a model [or system] should be disclosed to enable appropriate understanding,”
  - *Liao and Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap.*

# Measurement Validity

# Measurement Validity

- *Construct Reliability:* Is the construct itself valid?
- *Construct Validity*
  - *Face validity:* Does it intuitively make sense at all?
  - *Content validity:*
    - *Constestation:* Is there contestation around the construct?
    - *Substantive validity:* Are **only** relevant properties used?
    - *Structural Validity: Is there a relationship between properties and the construct?*

# Measurement Validity

- *Convergent Validity*
  - Is there a relation with existing measurements?
- *Discriminant Validity:*
  - Are some parts of the operationalisation shared w/ other constructs ?
- *Predictive Validity:*
  - Are the results from a measurement model predictive of its constituent constructs?

# Measurement Validity

- *Hypothesis Validity*
  - Can we draw interesting and meaningful hypotheses from the outcomes of the measurement model?
- *Consequential Validity*

# Construct Validity

- Consequential Validity
- Predictive Validity
- Hypothesis validity

# Actionability and Interpretability

- Interpretability as a field seeks to examine the process of arriving at a particular output
- Actionability asks whether we have enough information provided with the output to take any concrete steps

# Question Time

# An? Epistemology of Fairness?

- A mathematical epistemology of fairness
- What does a mathematical epistemology of fairness imply?

# An? Epistemology of Fairness?

- A mathematical epistemology of fairness
- What does a mathematical epistemology of fairness imply?
  - Only things that are countable can be understood to impact fairness
- The outcome here is that we can do things like operationalize utilitarianism

# A Motivation for Fairness?

- Blodgett et al., (2020) presents a strong motivation for clarifying what it means to be biased, and subsequently, what it means to be fair

# The trouble with defining “bias”

- The limitations of quantifiability of ”bias”
- “bias” is necessarily context dependent
- And is only meaningful in the context of marginalisation
- ML is for minimising the expectation of error
- And is built on human, political data
- Debiasing as a political act

# Defining “bias”

***Def 1:*** *Bias is the existence of an undesirable position with some imagination of a desirable position.*

# Defining “bias”

***Def 1:*** *Bias is the existence of an undesirable position with some imagination of a desirable position.*

***Def 2:*** *Bias is the systematic undesirable position produced with regard to existing systems of oppression.*

# What it means to be fair

- Often stated very clearly
- Different definitions of fairness may be incompatible
- As we saw in the readings with the recidivism example.
    - E.g., Predictive parity and equal opportunity

# Fairness definitions

- Individual Fairness
  - About ensuring that
- Group Fairness
- Fairness through unawareness

# Equalized Odds

- Equalized Odds
  - For all values $y \in Y, a \in A$

$$P(\hat{Y} = y \mid A = a, Y = y) = P(\hat{Y} = y \mid A = a', Y = y)$$

P: Probability

$\hat{Y}$: The predictions

Y: The ground truth (labelled data)

A: Protected characteristics

# Loan Example

Protected Attribute: $A$

- We can give out 10 loans
- We have 100 applicants
  - 70 come from affluent backgrounds
  - 30 come from low-income backgrounds
- Y = {Granted, Rejected}
- 50% of candidates from both groups are bad candidates

# Loan Example

| | Predicted Values | |
|---|---|---|
| Actual Values | True Positives | True Negatives |
| | False Positives | False Negatives |

# Loan Example

$$\text{Qualified applicants with affluent background} = 35$$
$$\text{Qualified applicants with low-income background} = 15$$

$$FPR_{affluent} = \frac{0}{35} = 0.0 \qquad TPR_{affluent} = \frac{7}{35} = 0.20$$

$$FPR_{low-income} = \frac{0}{15} = 0.0 \qquad TPR_{low-income} = \frac{3}{15} = 0.20$$

# Question Time

# Closing remarks

- Central to fairness and ethics
  - Be precise
  - Make sure that your methods and data fit together and can answer the same questions
  - Take measurement validity
    - Particularly attend to hypothesis validity
- Consent is king

# What are some examples of CSS you’ve seen?

- What are some CSS applications you’ve seen?
- Or that you think would be interesting or cool?