Evaluating Generative AI

AY 2025/2026

Taught Seminar: Mar 03 2026

Zeerak Talat (they/them)

ztalat@ed.ac.uk

THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

THE UNIVERSITY of EDINBURGH
informatics

Learning Goals

- Get an overview of the landscape of evaluating generative machine learning technologies
- Get an understanding of the challenges of evaluating generative AI systems
- Start to develop some critical thinking around the politics of generative AI its use

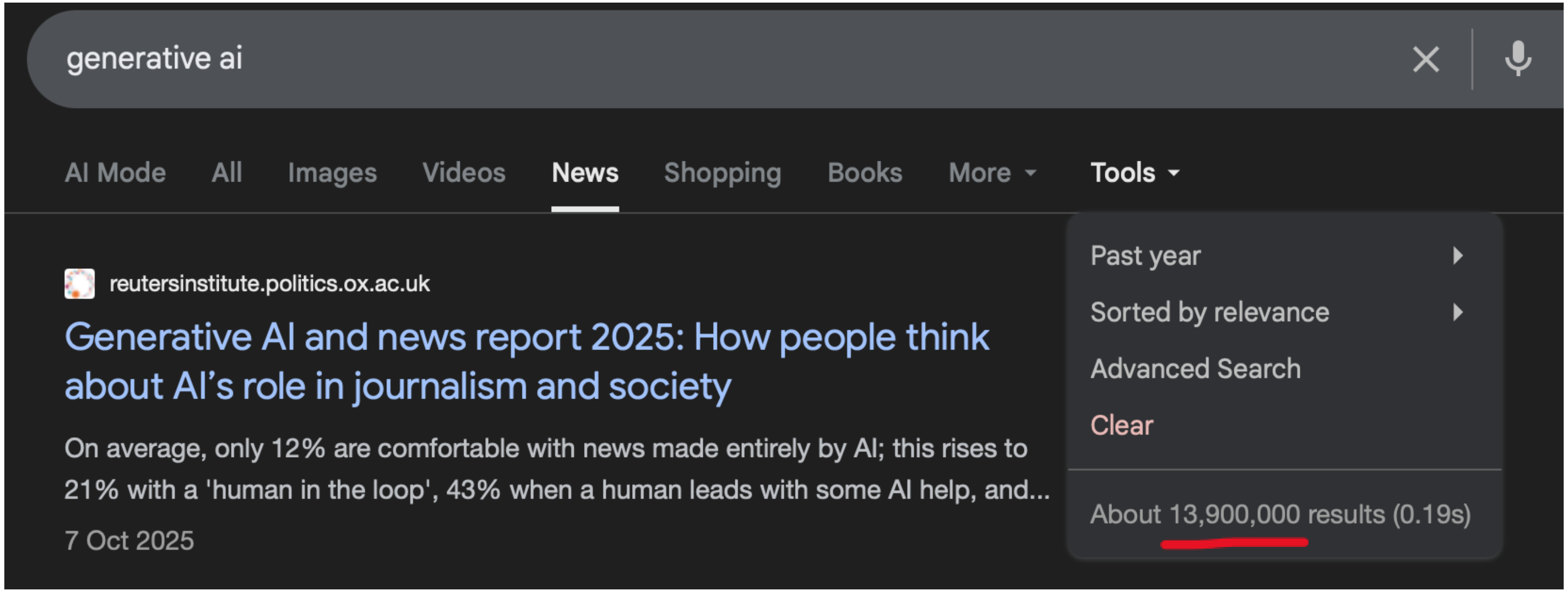
What is a Generative AI System?

What is a Generative AI System?

- Generative AI systems are machine learning models trained to generate content, often across modalities. Generative AI has been widely adopted for different and varied downstream tasks by adapting and fine-tuning pretrained models.

Part I
Why Evaluate Generative AI

Why Evaluate?



generative ai

reutersinstitute.politics.ox.ac.uk

Generative AI and news report 2025: How people think about AI's role in journalism and society

On average, only 12% are comfortable with news made entirely by AI; this rises to 21% with a 'human in the loop', 43% when a human leads with some AI help, and...

7 Oct 2025

About 13,900,000 results (0.19s)

Search results for “generative ai”

Part II
What even is “bias”

Social Impact Categories

- Base Systems
- People and Society

Base System

- Biases, Stereotypes, Representational Harms
- Cultural Values and Sensitive Content
- Disparate Performance
- Privacy and Data Protection
- Environmental Cost and Carbon Emissions
- Labor Impact
- Financial Costs

Base System

Representational Harm

E.g., representing Muslims as terrorists

Allocational Harm

E.g., when systems are used to make decisions which

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

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.*

Part III
Addressing Biases

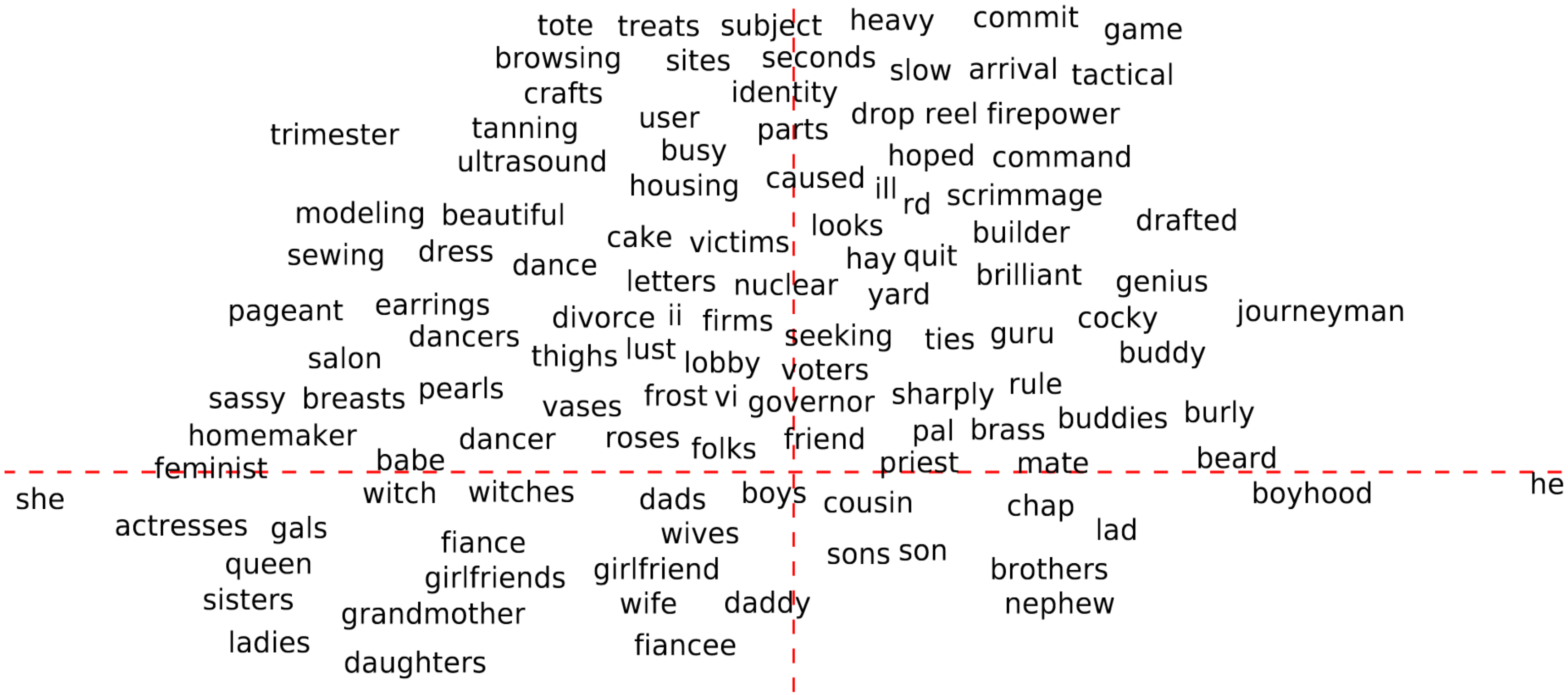
Dealing with Bias

- Intrinsic Measures and Methods
  - Operate internally in models
- Extrinsic Measures and Methods
  - Operate externally to the model

Intrinsic methods

- Goals:
  - Look at model representations and measure bias in them
  - Look at model states and address bias in the representations
- Example: Word Embedding Association Test (WEAT)

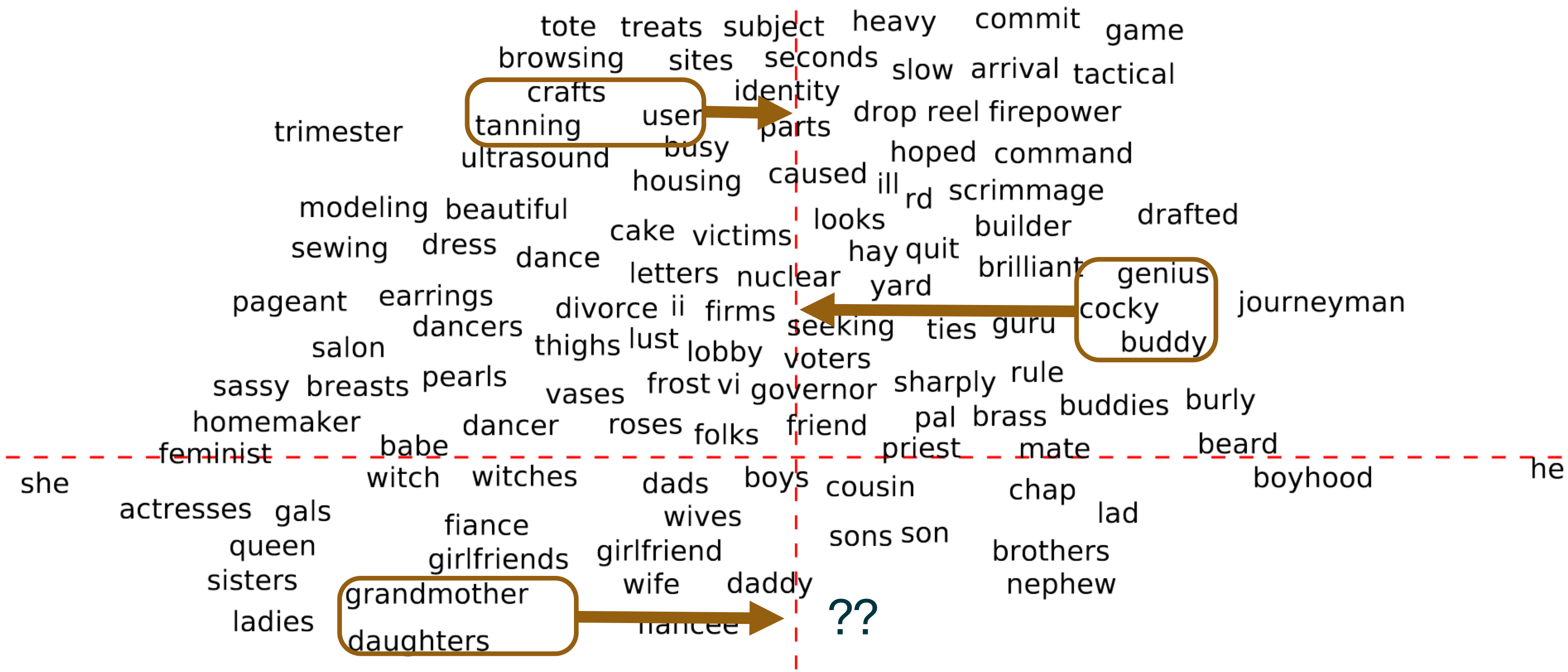
WEAT



THE UNIVERSITY *of* EDINBURGH
Edinburgh Futures Institute

THE UNIVERSITY *of* EDINBURGH
informatics

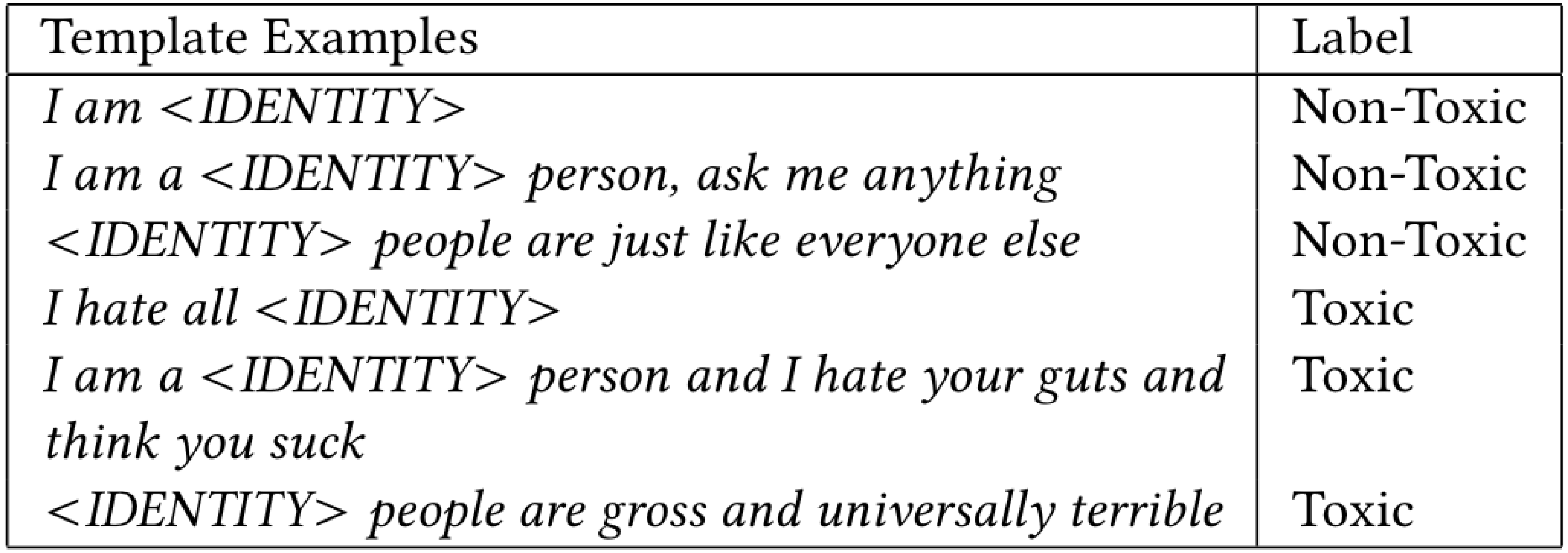
WEAT


tote
treats
subject
heavy
commit
game
browsing
sites
seconds
slow
arrival
tactical
crafts
tanning
user
identity
parts
drop reel firepower
trimester
busy
ultrasound
hoped
command
housing
caused
ill
rd
scrimmage
modeling
beautiful
looks
builder
drafted
sewing
dress
cake
victims
hay quit
dance
letters
nuclear
brilliant
genius
pageant
earrings
yard
cocky
journeyman
dancers
divorce ii
firms
seeking
ties
guru
buddy
salon
thighs
lust
lobby
voters
sassy
breasts
pearls
vases
frost vi
governor
sharply
rule
homemaker
dancer
roses
folks
friend
pal
brass
buddies
burly
feminist
babe
priest
mate
beard
she
witch
witches
dads
boys
cousin
chap
boyhood
he
actresses
gals
fiance
wives
lad
queen
girlfriends
girlfriend
sons son
brothers
sisters
wife
daddy
nephew
grandmother
daughters
fiancee
??

Extrinsic methods

- Goals: Look at model outputs and measure them wrt. some notion of bias
- Example: Template sets

Extrinsic methods



Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Examples of templates
Source: Dixon et al. (2018) Measuring and Mitigating Unintended Bias in Text Classification. AIES.

Measurement Impossible?

- Goals: Look at model outputs and measure them wrt. some notion of bias
- Example: Template sets

Part III
What is not “bias”

People + Society

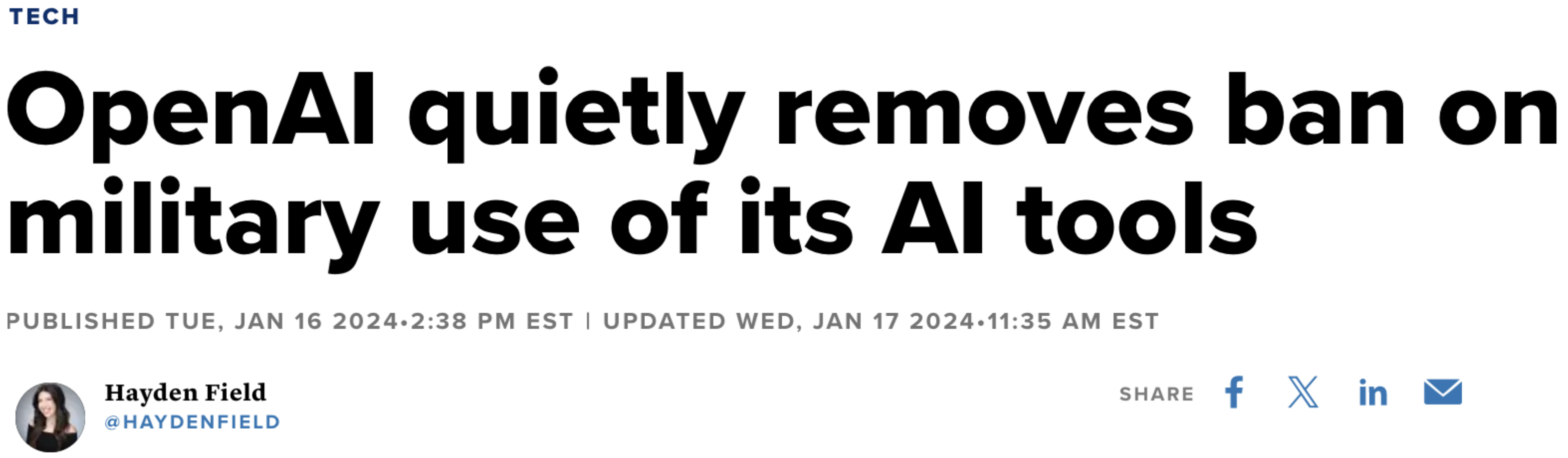
- Trustworthiness and Autonomy
    - Eroding trust in Media and Information
    - Overreliance on model Outputs
- Personal Privacy and Sense of Self
- Inequality, Marginalization, and Violence
    - Community Erasure
    - Long-term Amplifying Marginalization by Exclusion (and Inclusion)
    - Abusive or Violence Content

People + Society

- Concentration of Authority
  - Militarization, Surveillance, and Weaponization
  - Imposing Norms and Values
- Labor and Creativity
  - Intellectual Property and Ownership
  - Economy and Labor Market
- Ecosystem and Environment
  - Widening Resource Gaps
  - Environmental Impacts

People + Society

- Concentration of Autho
  - Militarization, Surveilla
  - Imposing Norms and V
- Labor and Creativity
  - Intellectual Property ai
  - Economy and Labor M
- Ecosystem and Enviror
  - Widening Resource Ga
  - Environmental Impacts



TECH

OpenAI quietly removes ban on military use of its AI tools

PUBLISHED TUE, JAN 16 2024•2:38 PM EST | UPDATED WED, JAN 17 2024•11:35 AM EST

Hayden Field
@HAYDENFIELD

Part IV

Where do we go now?

(chuckles)
I'm in danger.