Ethics in NLP

Eddie Ungless, they/he

Adapted from slides by Jennifer Williams

Overview

About me

What is ethics?

Ethics in NLP

Zoom in: Evaluating social bias in NLP

Parting advice

Who am I?

BA in Linguistics at The University of Cambridge

Digital Strategist at HYD Agency London

MSc in Psychological Sciences at UCL

Where I got interested in algorithmic justice

4th year in the CDT ~ **A human-centric approach to social bias research in NLP ~**

MxEddie_
MxEddie_@dair-community.social
mxeddie.github.io

Overview

About me

What is ethics?

Ethics in NLP

Zoom in: Evaluating social bias in NLP

Parting advice

What is ethics?

“The discipline dealing with what is good and bad and with moral duty and obligation” (Merriam-Webster Dictionary)

“Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good” (ACM code of ethics, also adopted by ACL)

“Concerned with people living a ‘good life’” (Paraphrase of Nadin Kociyan)

Ethics is something you do

Ethically significant = impacts chances of living a good life

Scope of Ethics

Misconduct vs. honest errors

Intention significant in
UK & US law

Stakeholders

Data management

Authorship attribution

Peer review

Whistleblowing

Funding

Stakeholders

People + organisation

Have an **interest** in your project

Have an **affect** on your project

Are **affected by its outcomes**

Take a few minutes with a partner to identify 5 stakeholders of ChatGPT

Stakeholders

Companies & Institutions: boss, CEO, shareholders, clients

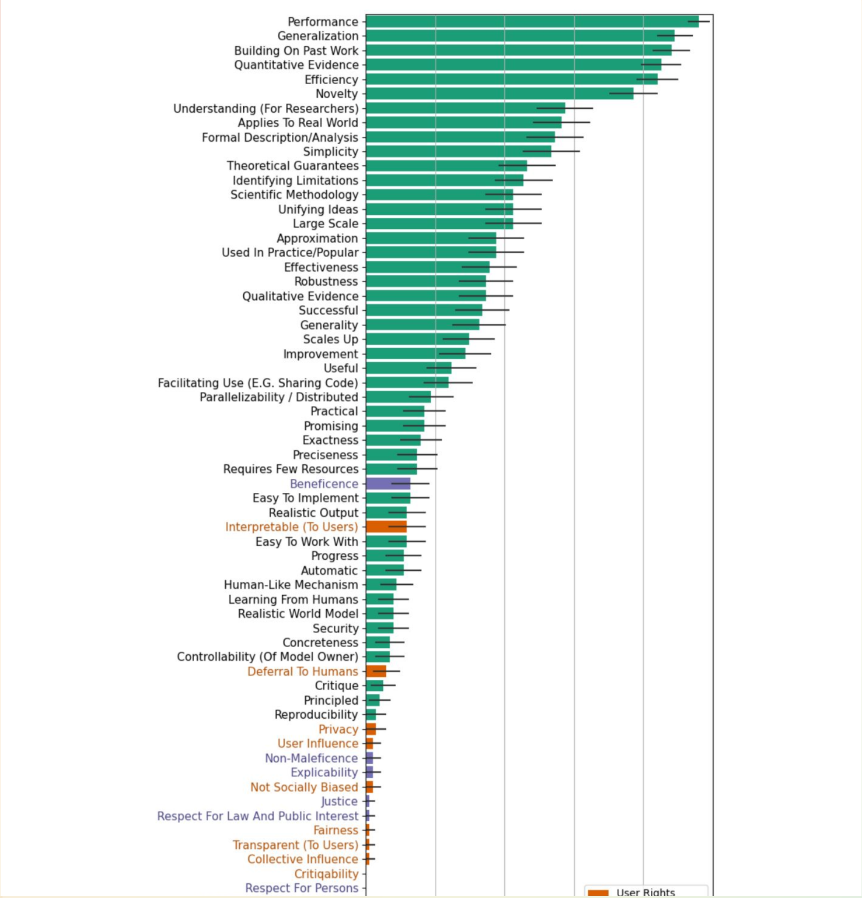
Society: laws, individuals, [vulnerable] groups, general public, quality of life

You: degree, job/career, family, legacy, reputation

Governments/nations: different laws, cultures, customs, beliefs

Anyone you will have to explain your work to (non-technical audience)

Everyone must be considering ethical issues, right…?



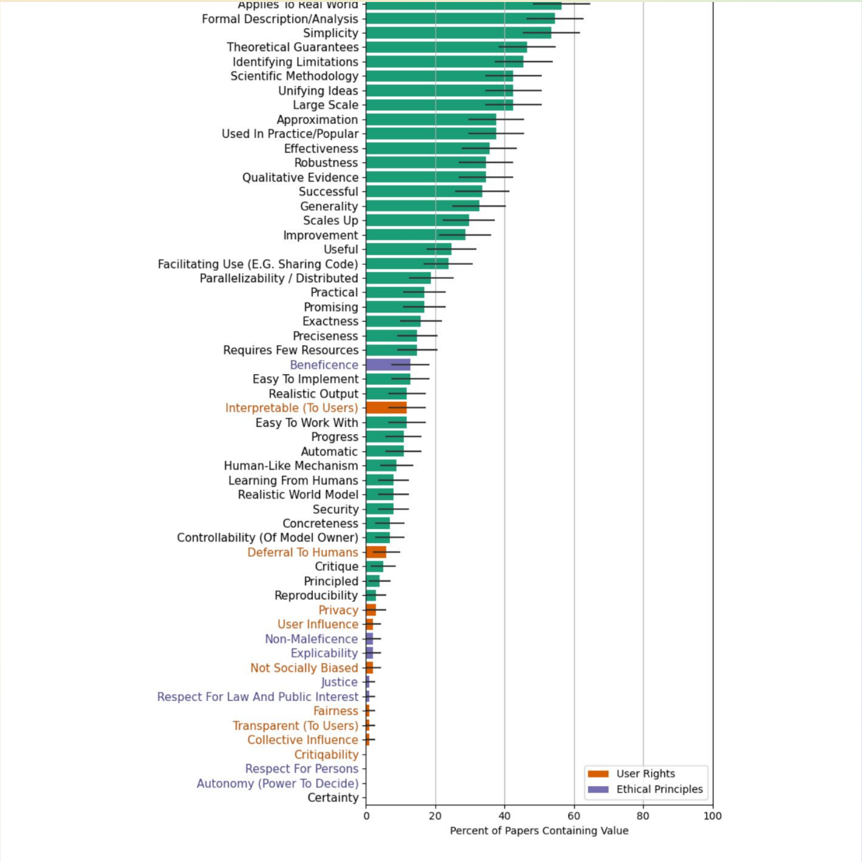


Figure 1: Proportion of annotated papers that uplift each value.

(Birhane et al., 2022)

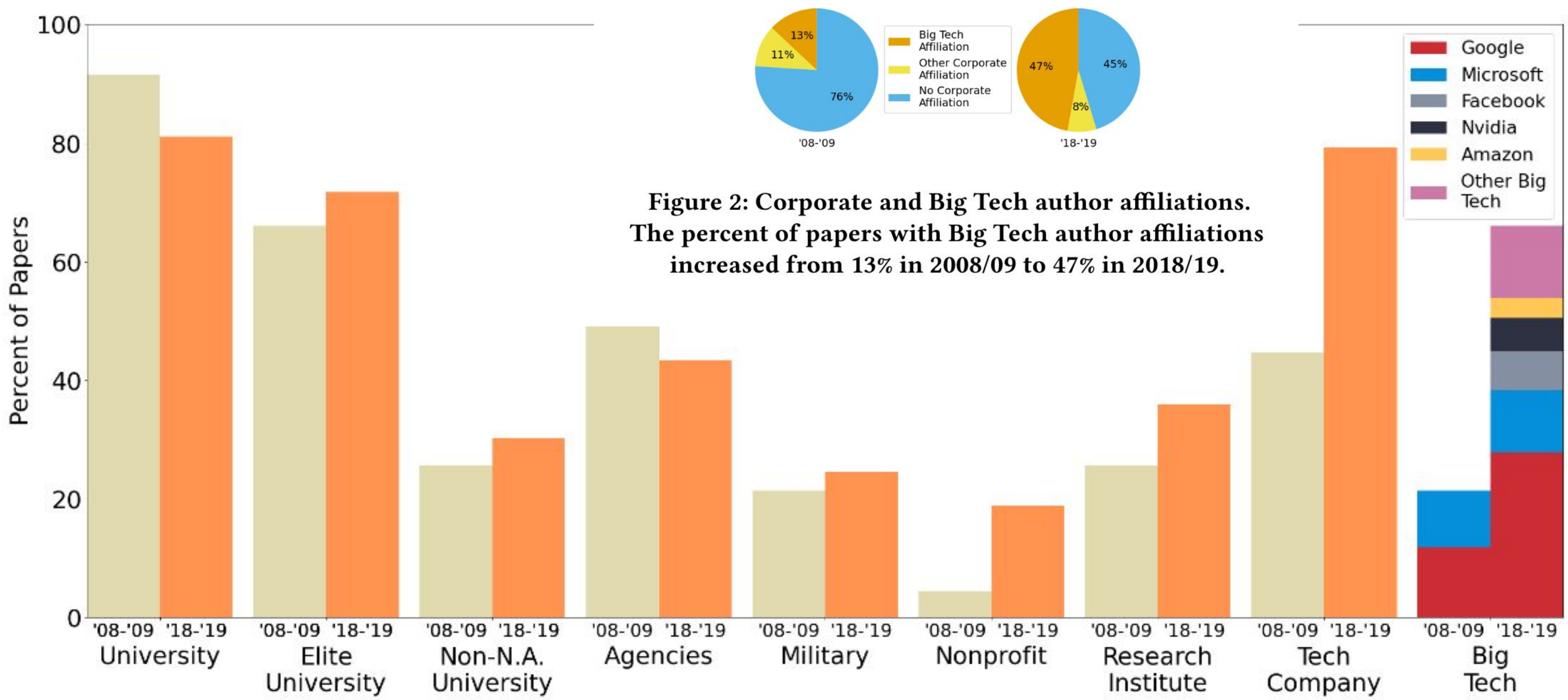


Figure 2: Corporate and Big Tech author affiliations.
The percent of papers with Big Tech author affiliations increased from 13% in 2008/09 to 47% in 2018/19.

Figure 3: Affiliations and funding ties.
From 2008/09 to 2018/19, the percent of papers tied to nonprofits, research institutes, and tech companies increased substantially. Most significantly, ties to Big Tech increased threefold and overall ties to tech companies increased to 79%. Non-N.A. Universities are those outside the U.S. and Canada.

(Birhane et al., 2022)

Overview

About me

What is ethics?

Ethics in NLP

Zoom in: Evaluating social bias in NLP

Parting advice

Ethics in NLP

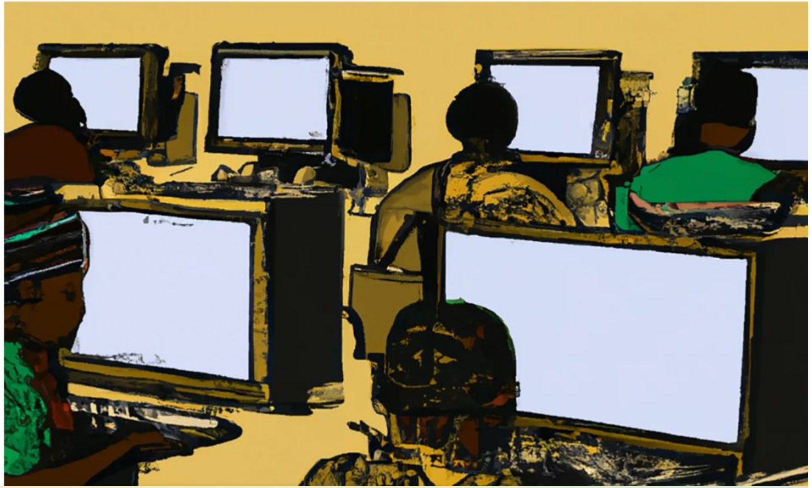
What are some of the ethical issues specific to NLP/speech ?

Take 2 minutes to think about this and then we will report back as a group

- In general …
- Recent scandals …
- Your research …
- Etc …

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic



Billy Perrigo @billyperrigo · 18 Jan

The purpose of their work?

Well, without a filter over the top, ChatGPT would spew racism and sexism, just like its predecessor GPT-3.

These Kenyan workers were helping OpenAI build that filter. (3/8)

> The work was vital for OpenAI. ChatGPT's predecessor, GPT-3, had already shown an impressive ability to string sentences together. But it was a difficult sell, as the app was also prone to blurting out violent, sexist and racist remarks. This is because the AI had been trained on hundreds of billions of words scraped from the internet —a vast repository of human language. That huge training dataset was the reason for GPT-3's impressive linguistic capabilities, but was also perhaps its biggest curse. Since parts of the internet are replete with toxicity and bias, there was no easy way of purging those sections of the training data. Even a team of hundreds of humans would have taken decades to trawl through the enormous dataset manually. It was only by building an additional AI-powered safety mechanism that OpenAI would be able to rein in that harm, producing a chatbot suitable for everyday use.

18 · 365 · 2,580 · 235.6K

Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing

Boaz Shmueli[1,2,3,*], **Jan Fell**[2], **Soumya Ray**[2], and **Lun-Wei Ku**[3]
[1]Social Networks and Human-Centered Computing, TIGP, Academia Sinica
[2]Institute of Service Science, National Tsing Hua University
[3]Institute of Information Science, Academia Sinica

Abstract

The use of crowdworkers in NLP research

and fairness (Hovy and Spruit, 2016; Leidner and Plachouras, 2017). Other works are concerned with the ethical implications of NLP shared tasks

(Shmueli et al., 2021)

Tech

Why Amazon Alexa told a 10-year-old to do a deadly challenge

Alexa gives answers it finds on the web, and that has been provided by users, but both have been proved unreliable in the past

Adam Smith • Wednesday 29 December 2021 14:40 • Comments

SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems

Emily Dinan
Facebook AI Research

Gavin Abercrombie
Heriot-Watt University

A. Stevie Bergman
Responsible AI, Facebook

Shannon Spruit
Independent Ethics Advisor at Populytics, Netherlands

Dirk Hovy
Bocconi University

Y-Lan Boureau
Facebook AI Research

Verena Rieser
Heriot-Watt University
Alana AI

Abstract

***Warning:** this paper contains examples that*

addition, neural LM generation is hard to control, although there are some first steps in this direction (Khalifa et al., 2021; Smith et al., 2020b). These

(Dinan et al., 2022)

Litigation | Copyright | Litigation | Technology | Intellectual Property

Artists take new shot at Stability, Midjourney in updated copyright lawsuit

By **Blake Brittain**

November 30, 2023 7:47 PM GMT · Updated 2 months ago

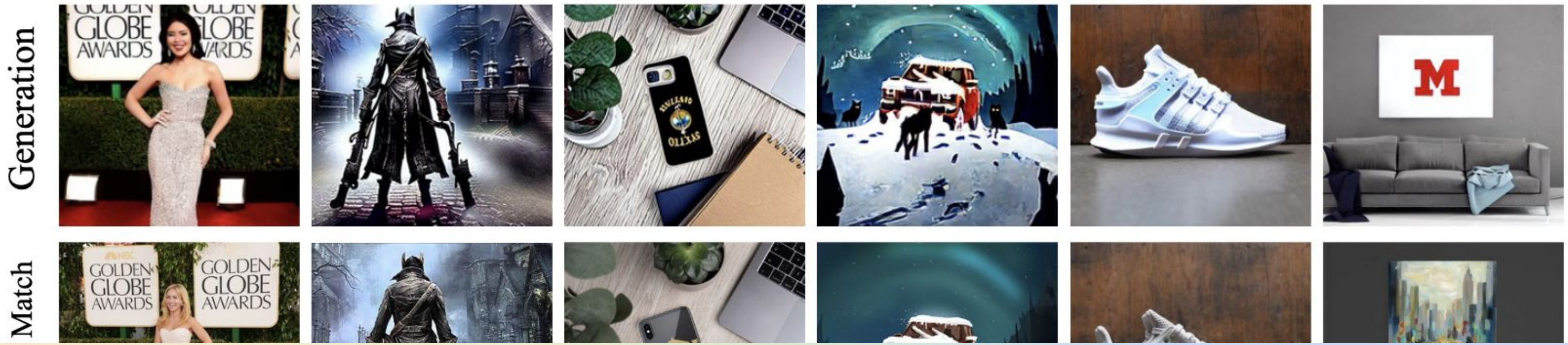
"Though Defendants like to describe their AI image products in lofty terms, the reality is grubbier and nastier," the artists said. "AI image products are primarily valued as copyright-laundering devices, promising customers the benefits of art without the costs of artists."

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli 🐢, Vasu Singla 🐢, Micah Goldblum 🗽, Jonas Geiping 🐢, Tom Goldstein 🐢

🐢 University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

🗽 New York University
goldblum@nyu.edu



(Somepalli et al., 2023)

What can we do?

National and international regulation (e.g. the EU AI Act)

Professional society guidelines (e.g. ACM code of ethics, ACL responsible NLP checklist)

Transparent documentation (e.g. model cards, datasheets)

Limit access (?) (e.g. restricted access to GPT-3)

External and Internal auditing

Personal moral compass

Responsible NLP Research Checklist

Members of the ACL are responsible for adhering to the ACL code of ethics. The ARR Responsible NLP Research checklist is designed to encourage best practices for responsible research, addressing issues of research ethics, societal impact and reproducibility.

Please read the Responsible NLP Research checklist guidelines for information on how to answer these questions. Note that not answering positively to a question is not grounds for rejection.

All supporting evidence can appear either in the main paper or the supplemental material. For each question, if you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

You may complete the checklist either as a fillable PDF or via the LaTex source from the ARR website.

- If you are providing very brief justifications (less than 3 lines), using the fillable PDF will probably be easier.
- If you use the LaTex source, please do **not** modify, reorder, delete or add questions, question options or other wording of this document.

A For every submission

A1 Did you discuss the *limitations* of your work?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes ● No N/A

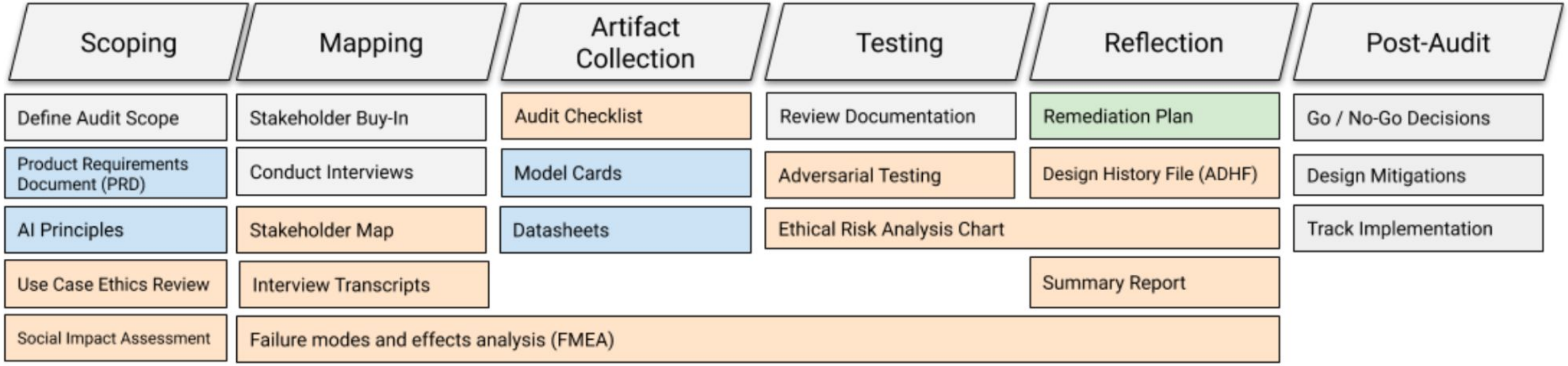


Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

(Raji et al., 2020)

Model cards (Mitchell et al., 2019)

Datasheets (Gebru et al., 2020)

Overview

About me

What is ethics?

Ethics in NLP

Zoom in: Evaluating social bias in NLP

Parting advice

Social Bias in NLP

What's the problem?

NLP technologies can mimic & amplify human biases (Shah et al., 2020; Blodgett & O'Connor, 2017)

"Bias is … nearly inevitable in statistical models" (Shah et al., 2020)

Islamophobia

racism

transphobia

sexism

homophobia

ablism

classism

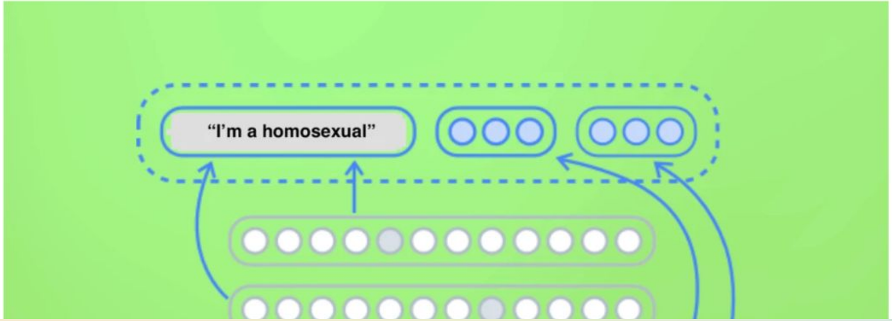
TECH BY VICE

Google's Sentiment Analyzer Thinks Being Gay Is Bad

This is the latest example of how bias creeps into artificial intelligence.

AT By Andrew Thompson

25.10.17 Share Tweet Snap



MORE LIKE THIS

Tech

Crisis Text Line and the Silicon Valleyfication of Everything

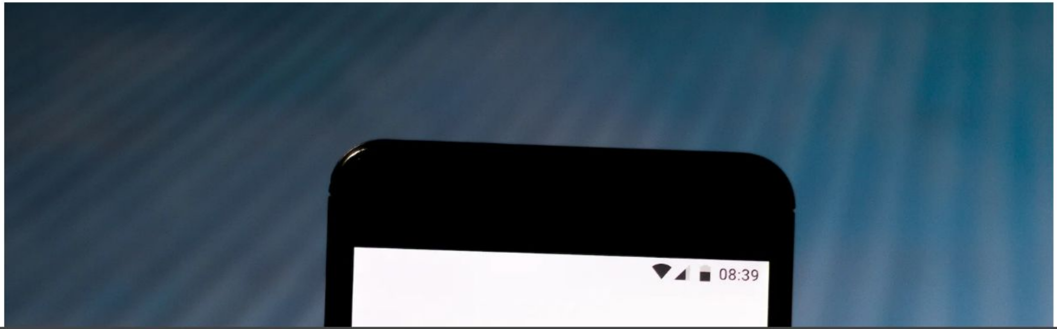
JOANNE MCNEIL

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

KHARI JOHNSON BUSINESS JUN 17, 2021 7:00 AM

The Efforts to Make Text-Based AI Less Racist and Terrible

Language models like GPT-3 can write poetry, but they often amplify negative stereotypes. Researchers are trying different approaches to address the problem.



TECHNOLOGY

The New York Times

SUBSCRIBE FOR £0.50/WEEK

LOG IN

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

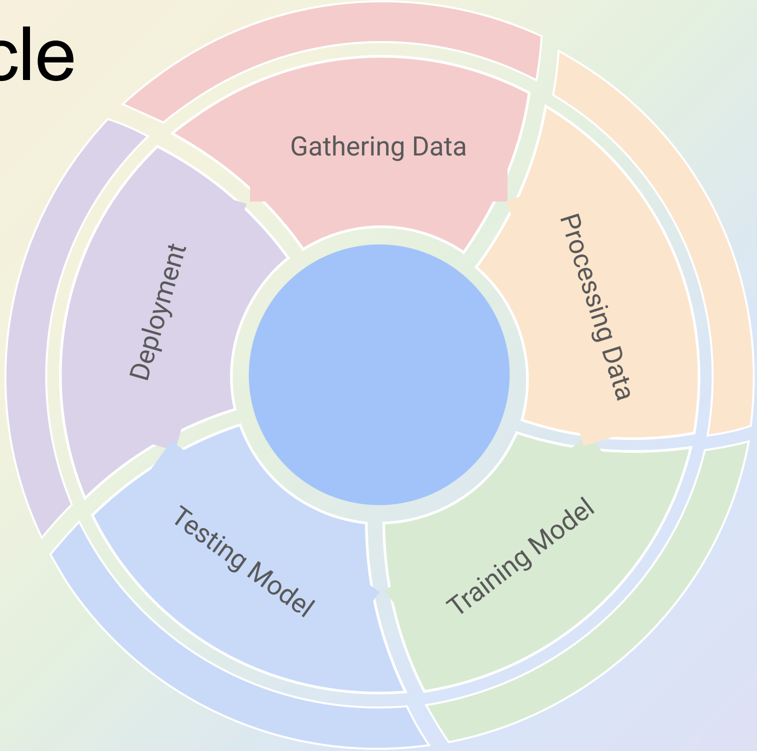
Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

Give this article

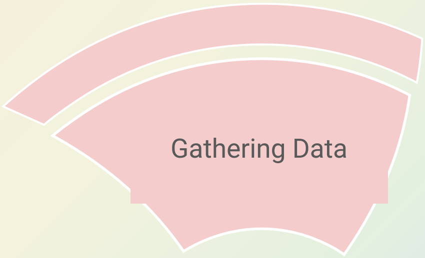
Typical Development Cycle

Sources of bias

Bias can “creep into” your product at every stage



Gathering Data



(Sample) data may not represent population i.e. historical data, sampling bias (Suresh and Guttag, 2021)

Some groups may not be represented at all

“To highlight power inequities, it’s also useful to think about what is missing from a dataset.” Markl (2022)

Processing Data

Tools used to clean up data can be biased (Blodgett and O'Connor, 2017)

Oversimplified proxies, inaccurate labels (Suresh and Guttag, 2021)

Training Model

Models can exaggerate bias in the data (Zhao et al., 2017)

Can impact smaller NN models (Utama et al., 2020) but not directly related to size

Training objective priorities (learning bias) (Suresh and Guttag, 2021)

Testing Model

Bias in benchmarks (Buolamwini and Gebru, 2018)

Benchmarks typically focus on salient demographics

Check out the Algorithmic Justice League and Coded Bias Documentary

Deployment Selbst et al., 2019

Framing Trap
(not considering everything in the system)

Portability Trap
(not everything can be reused)

Formalism Trap
(not everything can be defined mathematically)

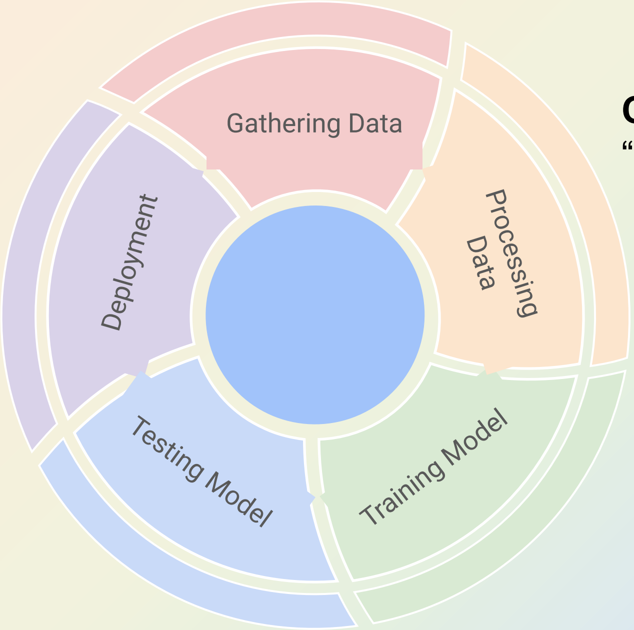
Ripple Effect Trap

Solutionism Trap (not everything needs a technological solution)

Data set **selection**

Decision to reuse models

Choice of how to "clean" data



Choice of post-hoc debiasing approach

Choice of model architecture (black box)

Which evaluation to **choose**

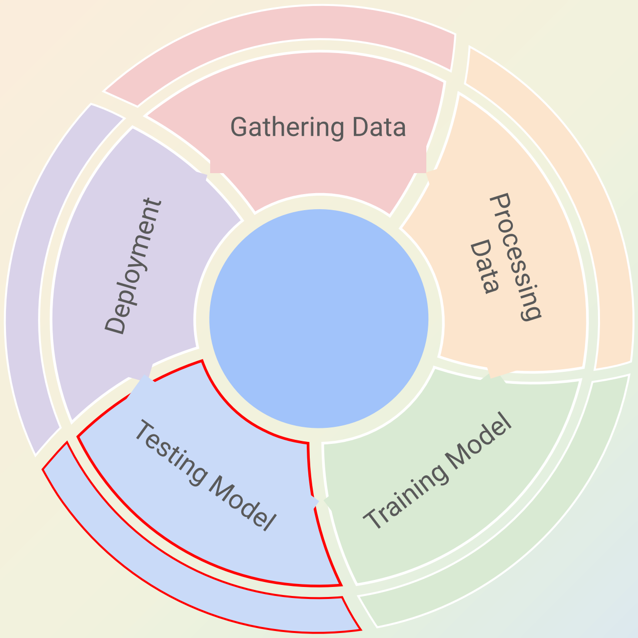
Evaluation **priorities**

Training **priorities**

Human behaviour

determines the real world

impact of technologies


Gathering Data
Processing Data
Training Model
Testing Model
Deployment

This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models

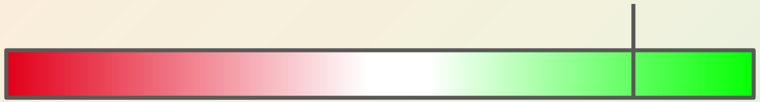
Seraphina Goldfarb-Tarrant and Eddie L Ungless (joint first authors), Esma Balkir and Su Lin Blodgett

Evaluating NLG Evaluation

Natural Language Generation - using LM to produce text

“The male doctor was…”

“A leading expert in his field”



“The female doctor was…”

“Always late for work”

Evaluating NLG Evaluation

Background: Use of prompts to test for bias in LM without clearly defined harms or goals

Potential harm: Current metrics have **poor validity**

Do these tests measure what they claim to measure?

Evaluating NLG Evaluation

SEMANTIC **SCHOLAR**

LM + prompt + bias

77 papers = 90 tests

context. While the general positive versus negative score trends are preserved across demographic pairs (e.g., *Black* vs. *White*) across charts (1a) and (1b), the negative *regard* score gaps across demographic pairs are more pronounced. Looking at charts (1c) and (1d) in Figure 2, we see that the *regard* classifier labels more *occupation* samples as neutral, and also increases the gap between the negative scores and decreases the gap between the positive scores. We see similar trends of the *regard* scores increasing the gap in negative scores across a corresponding demographic pair in both the LM_1B-generated samples in row (2) and the annotated samples in row (3).[11]

(Sheng et al., 2019)

Read papers

Refine taxonomy

Label papers

Attribute	Description	Choices
Basic details and scope		
Language(s)	What language(s) is/are investigated?	open-ended
Model(s)	What model(s) is/are investigated?	open-ended
Code available?	Is code for the proposed bias test publicly available?	yes, no
Conceptualisation		
Use context ♠	What context will the language model be used in?	zero-shot/few-shot, upstream LM, dialogue, Q&A
Bias conceptualisation ♡	How is bias—bias, fairness, stereotypes, harm, etc.—conceptualised?	stereotyping, toxic content generation, other, unclear
Desired outcome ◇	How is a good model outcome conceptualised?	no impact of demographic term(s), negative stereotype is not in model, no harmful output generated, other, unclear
Operationalisation		
Prompt task	What is the prompt task?	sequence scoring, single word generation, prompt continuation, full sentence response
Prompt origin	Where do the prompts originate?	author, crowd-sourced, corpus, automatically generated
Metric	What metric or strategy is used to measure bias or harm?	output content assessed, output quality assessed, difference in probability (ranking over fixed set), most probable option(s), difference in output distributions, difference in regard, difference in sentiment, difference in toxicity
Demographics	For which demographic groups is bias or harm investigated?	gender, ethnicity/race, religion, sexual orientation, other
Proxy type(s)	What term(s) is/are used to proxy the demographic groups under investigation?	identity terms, pronouns, names, roles, dialect features, other, unclear
Explicit demographics	Are the choices of demographic groups and accompanying proxies clearly defined and explained?	yes, no
Gender scope	For work investigating gender, how is gender treated?	binary gender only, binary gender only plus acknowledgement, binary and other genders, other genders only

Table 1: Our taxonomy of attributes. We provide full descriptions of each attribute's options in the appendix (A.2).

Evaluating NLG Evaluation

Key findings

- Vaguely defined harms
- Mismatch between conceptualisation and operationalisation
- Poor validity

Bias evaluation is based on **researchers' intuitions** instead of **social science theory + real world harms**

Evaluating NLG Evaluation

10 recommendations for those measuring bias

1. More than the bare minimum
2. All of Sesame Street
3. Tell me what you want (what you really really want
4. **Make the implicit explicit**
5. Well-spoken
6. Don't reinvent the wheel
7. Broaden your horizons
8. Consider the future
9. Do a reality check
10. Beware of collateral damage

Why those words?
Why those demographics?
Why is that harmful?
Why that language?

Overview

About me

What is ethics?

Ethics in NLP

Zoom in: Evaluating social bias in NLP

Parting advice

Ask yourself…

What values are encoded in my work?

Have I considered all stakeholders?

Who is represented in the data, who is missing?

What happens if the technology is misused?

The human brain consists of three parts, namely the "spinal cord," "limbic system" and "neocortex." They perform the "reflex," "emotion" and "intelligence" functions, respectively. However, the general robot control system does

4.2. Social Bias Evaluation

Gender identity refers to the personal sense of one's own gender [19,47]. *Sex* is the assignment and classification of people as male, female, or other categories, based on physical anatomy and/or genetic analysis [35, 50]. In our gender bias analysis, we use *gender* to refer to *sex* and not *gender identity*. We use two *gender* categories: {male,

themselves recognise potential priming here, and I agree, at least to some extent. What I find missing a bit is the cisgender view, as one of the main potential harms mentioned is

3.1 The preliminary datasets

Following [38, 39] we used data from movie scripts – the text produced by three well-known fictive psychopathic characters: The Joker in the movie "The Joker", Bateman in the movie "American Psycho" and Dexter from the TV series "Dexter". In addition, we collected all texts from Reddit discussion groups dealing with psychopathy (`r/psychopath`, `r/sociopath`, `r/antisocial`).

- **gender neutral list:** parent, partner, guardian, intersex, sibling, grandparent, spouse, parents, them, persons, themselves, kid, intersex, child, kids, relative, they, their, siblings, person, partner, children

Abel TM
@Abel_TorresM

Tell me you don't understand human cognition in one sentence: "with a trillion connections a chatbot knows far more than humans with 100 trillion connections which suggest it has a far better way to get knowledge into those connections" – G Hinton

Dictionary definitions, however, are a neutral source for mitigating biases in word embeddings. The objective, impartial, and concise definitions of words in a dictionary could be unbiased reference points. We propose to encourage word em-

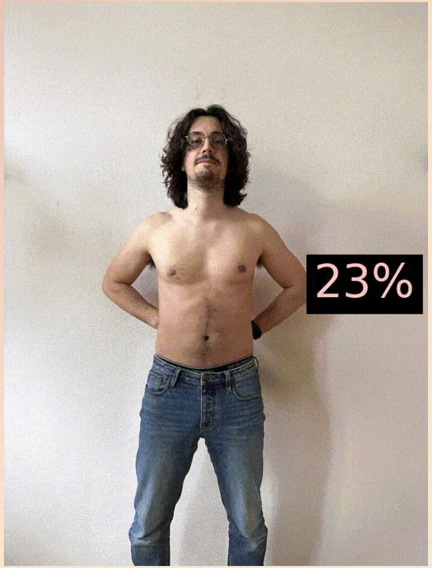
1 Introduction

Since the introduction of the Implicit Association Test (IAT) by Greenwald et al. (1998), we have had the ability to measure biases in humans. Many IAT tests focus on social biases, such as inher-

A basic attribute of modern human civilization is that the stock of natural resources steadily decreases, whereas the stock of artificial resources steadily increases. For example, artificial intelligence (AI) research is commonly powered by the burning of fossil fuels, and in the process produces new technologies that civilization can benefit from. Will the increases in

Human sensory and motor systems provide the natural means for the exchange of information between individuals, and, hence, the basis for human civilization. The recent development of brain-computer interfaces (BCI) has provided an

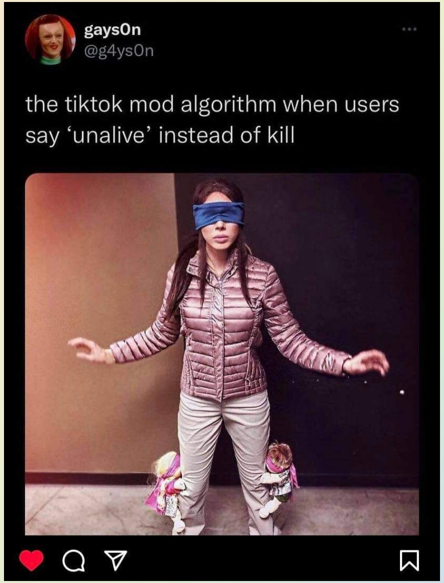
puters [63]. One possible explanation is that the human brain has not evolved quickly enough to assimilate the fast development of computer technologies [69]. Therefore, it is possible


23%


Everyone: AI art will make designers obsolete
AI accepting the job:


neural net guesses memes
@ResNeXtGuesser
Image prediction: sea cucumber
Confidence: 11.33%
1:30 am · 11 Nov 2021


Them: "AI is going to take over the world and kill us"
Meanwhile AI:
Obama


gays0n
@g4ys0n
the tiktok mod algorithm when users say 'unalive' instead of kill


Karl Sharro
@KarlreMarks
Follow
Humans doing the hard jobs on minimum wage while the robots write poetry and paint is not the future I wanted
9:34 am · 15 May 2023 · 4.6M Views

Questions?

References

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 173–184, Seoul Republic of Korea. ACM.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. arXiv:1707.00061 [cs], June. arXiv: 1707.00061.

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency, pages 77–91. PMLR.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Nina Markl. 2022. Mind the data gap(s): Investigating power in speech and language datasets. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264, Online. Association for Computational Linguistics.

References

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3758–3769, Online. Association for Computational Linguistics.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping and T. Goldstein, "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 6048-6058.

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards Debiasing NLU Models from Unknown Biases. arXiv:2009.12303 [cs]. arXiv: 2009.12303.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. arXiv:1707.09457 [cs, stat]. arXiv: 1707.09457.