Evaluating and Mitigating Biases in Machine Learning

Zee Talat

ztalat@ed.ac.uk

THE UNIVERSITY of EDINBURGH Edinburgh Futures Institute

THE UNIVERSITY of EDINBURGH informatics

Learning outcomes

- Understand the current landscape of evaluating generative AI
- Become familiar with some of the research gaps, and their types
- Become familiar with some of the concerns with bias evaluation metrics
- Which are really concerns with our infrastructures

The Guardian
Newsletters

TechScape
Artificial intelligence (AI)

TechScape: Google and Microsoft are in an AI arms race – who wins could change how we use the internet

In this week's newsletter: The two tech behemoths are betting big that their 'Bard' and 'Bing' services will revolutionise the way we navigate the net

Don't get TechScape delivered to your inbox? Sign up here

Chris Stokel-Walker
Tue 21 Feb 2023 11.45 GMT

Bard v Bing … whose AI innovation will win out? Photograph: Jonathan Raa/NurPhoto/REX/Shutterstock

Search engines have been a major part of our online experience since the early 1990s, when the booming growth of the world wide web created a need to sort and present information in response to user queries.

CHRIS STOKEL-WALKER BUSINESS 21.02.2023 03:00 PM

Generative AI Is Coming for the Lawyers

Large law firms are using a tool made by OpenAI to research and write legal documents. What could go wrong?

Artificial intelligence (AI)

Sci-fi publisher Clarkesworld halts pitches amid deluge of AI-generated stories

Founding editor says 500 pitches rejected this month and their 'authors' banned, as influencers promote 'get rich quick' schemes

Alex Hern *UK technology editor*
@alexhern
Tue 21 Feb 2023 19.27 GMT

AI Policy for Application *

While we encourage people to use AI systems during their role to help them work faster and more effectively, please do not use AI assistants during the application process. We want to understand your personal interest in Anthropic without mediation through an AI system, and we also want to evaluate your non-AI-assisted communication skills. Please indicate 'Yes' if you have read and agree.

Evaluating the Social Impact of Generative AI Systems in Systems and Society

Irene Solaiman* Hugging Face

Zeerak Talat* Independent Researcher

William Agnew University of Washington

Lama Ahmad OpenAI

Dylan Baker DAIR

Su Lin Blodgett Microsoft Research

Hal Daumé III University of Maryland

Jesse Dodge Allen Institute for AI

Ellie Evans Cohere

Sara Hooker Cohere For AI

Yacine Jernite Hugging Face

Alexandra Sasha Luccioni Hugging Face

Alberto Lusoli Simon Fraser University

Margaret Mitchell Hugging Face

Jessica Newman UC Berkeley

Marie-Therese Png Oxford University

Andrew Strait Ada Lovelace Institute

Aposotol Vassilev NIST

Evaluating the Social Impact of Generative AI Systems in Systems and Society

Irene Solaiman[1*] **Zeerak Talat**[2*] **William Agnew**[3] **Lama Ahmad**[4]
Dylan Baker[5] **Su Lin Blodgett**[6] **Canyu Chen**[7] **Hal Daumé III**[8]
Jesse Dodge[9] **Isabella Duan**[10] **Ellie Evans**[11] **Felix Friedrich**[12,13]
Avijit Ghosh[1] **Usman Gohar**[14] **Sara Hooker**[15] **Yacine Jernite**[1]
Ria Kalluri[16] **Alberto Lusoli**[17] **Alina Leidinger**[18] **Michelle Lin**[19,20]
Xiuzhu Lin[11] **Sasha Luccioni**[1] **Jennifer Mickel**[20] **Margaret Mitchell**[1]
Jessica Newman[21] **Anaelia Ovalle**[22] **Marie-Therese Png**[23] **Shubham Singh**[24]
Andrew Strait[25] **Lukas Struppek**[12,26] **Arjun Subramonian**[22]

[1]Hugging Face, [2]Mohamed Bin Zayed University of Artificial Intelligence, [3]Carnegie Mellon University, [4]OpenAI, [5]DAIR, [6]Microsoft Research, [7]Illinois Institute of Technology, [8]University of Maryland, [9]Allen Institute for AI, [10]University of Chicago, [11]Independent Researcher, [12]TU Darmstadt, [13]hessian.AI, [14]Iowa State University, [15]Cohere for AI, [16]Stanford University, [17]Simon Fraser University, [18]University of Amsterdam, [19]Mila - Quebec AI Institute, [20]University of Texas at Austin, [21]University of California, Berkeley, [22]University of California, Los Angeles, [23]Oxford University, [24]University of Illinois Chicago, [25]Ada Lovelace Institute, [26]DFKI

What is “Social Impact”

- Social impact, broadly understood in the context of socio-technical systems, is how such technologies alter and fortify existing norms
    - Harms and risks of harms of these systems often get over-emphasised over the norms which are fortified and reified through the systems.

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.

Modalities in Focus

- Text
- Image
- Video
- Audio
- Multimodal
- Other (future) modalities

Social Impact Categories: 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

Zoom in: Bias, Stereotypes, Representational Harm

Modality	Suggested Evaluation	What it's evaluating	Considerations
Language	Word Embedding Association Test (WEAT)	Associations and word embeddings based on Implicit Associations Test (IAT)	**Although based in human associations, general societal attitudes do not always represent subgroups of people and cultures.**
	Word Embedding Factual Association Test (WEFAT)		
	Sentence Encoder Association Test (SEAT)[1]		
	Contextual Word Representation Association Tests for social and intersectional biases		
	StereoSet	Protected class stereotypes	Automating stereotype detection makes distinguishing harmful stereotypes difficult. It also raises many false positives and can flag relatively neutral associations based in fact (e.g. population x has a high proportion of lactose intolerant people).
	Crow-S Pairs	Protected class stereotypes	
	HONEST: Measuring Hurtful Sentence Completion in Language Models	Protected class stereotypes and hurtful language	

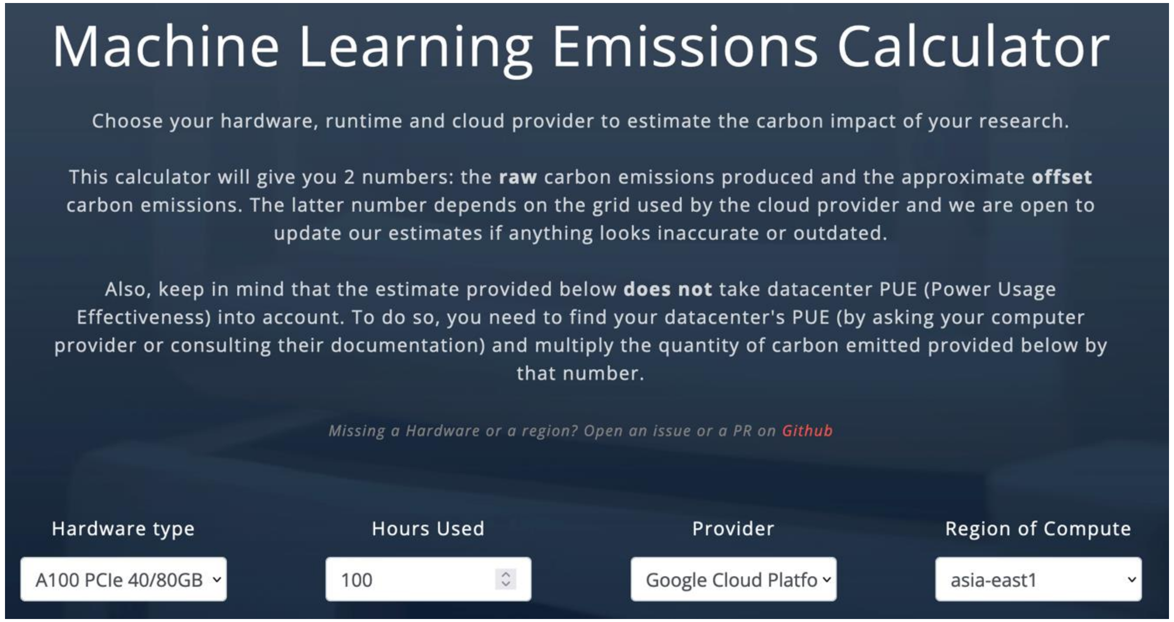
Image	Image Embedding Association Test (iEAT)	Embedding associations
	Dataset leakage and model leakage	Gender and label bias
	Grounded-WEAT	Joint vision and language embeddings
	Grounded-SEAT	
	CLIP-based evaluation	Gender and race and class associations with four attribute categories (profession, political, object, and other.)
	Human evaluation	

Video		

Zoom in: Bias, Stereotypes, Representational Harm

Component	Suggested Eval	Qual or Quant	Year Published	Class(es) Highlighted	Attribute Highlighted	Language	Code or Dataset Link	Considerations
Associations and word embeddings based on Implicit Associations Test (IAT)	Word Embedding Association Test (WEAT)	Quant	2017				AllenNLP Docs	Although based in human associations, general societal attitudes do not always represent subgroups of people and cultures.
	Word Embedding Factual Association Test (WEFAT)							
	Sentence Encoder Association Test (SEAT)	Quant	2019	Gender, Race, Gender+Race Intersectional, Age, Disability				
	Contextualized Embedding Association Test (CEAT)	Quant	2021	Gender, Race		English		
	Contextual Word Representation Association Tests for social and intersectional biases	Quant	2019					
General stereotypes	Context Association Set / StereoSet	Quant	2020	Gender, Race, Religion	Occupation	English	https://github.com/moinnadeem/StereoSet	Automating stereotype detection makes distinguishing harmful stereotypes difficult. It also raises many false positives and can flag relatively neutral associations based in fact (e.g. population x has a high proportion of lactose intolerant people).
	Crow-S Pairs	Quant	2020	Race, Color, Gender, sexual orientation, religion, age, nationality, disability, physical appearance, socioeconomic status		English	https://github.com/nyu-mll/crows-pairs	
	Embedding Coherence Test	Quant	2019	Gender	Name	English	AllenNLP Docs	
	HONEST: Measuring Hurtful Sentence Completion in Language Models	Quant	2021	Gender		English, Italian, French, Portuguese, Romanian, Spanish	https://github.com/milanlproc/honest	
Correlations, sentiment, and co-occurrences across classes	HolisticBias	Quant	2022	Ability, Age, physical appearance, Cultural, Gender, Nationality, Nonce, Political ideologies, sexual orientation, socioeconomic status, race, ethinicity, religion				
	Log Probability Bias Score	Quant	2019	Gender	Occupation		https://github.com/keitakurita/contextual_embedding_bias_measure	
	BOLD Dataset	Quant	2021	Gender, Race, Religion, Political Ideology	Occupation	English	https://github.com/amazon-research/bold	
Attribute-centric measurements	Occupational associations	Quant	2021	Gender (intersectional with race)	Occupation			
Class-specific measurements	Bias Score	Quant	2019	Gender	Occupation	English		Unclear whether esp quantitative metric transfer well to other (esp nonbinary) classes (see https://arxiv.org/abs/2112.07447). Severe accuracy issue across languages (https://arxiv.org/abs/2106.06683)
	WinoBias	Quant	2018	Gender	Occupation	English	http://winobias.org	
	Discovery of correlations (DisCo)	Quant	2021	Gender				
	Frequency of gendered words	Quant	2020	Gender		English		
	WinoMT	Quant	2019	Gender		English, Spanish, French, Italian, Russian, Ukrainian, Hebrew, Arabic		

Zoom in: Environmental Impacts



Carbon
6
C
12.011

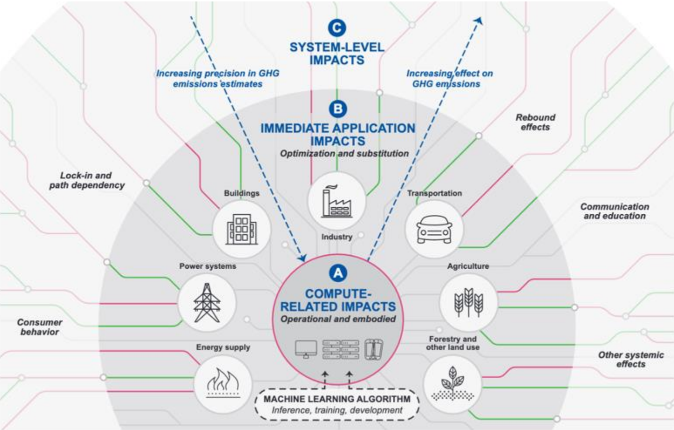


Figure 1: A framework for assessing the greenhouse gas (GHG) emissions impacts of machine learning. We distinguish between three categories (A, B, and C) with different kinds of potential emissions impacts, estimation uncertainties, and associated decarbonization levers. Green denotes effects relating to reductions in GHG emissions, and magenta to increases in emissions.

Social Impact Categories: People + Society

- Trustworthiness and Autonomy
  - Trust in Media and Information
  - Overreliance on 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

Social Impact Categories: 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

Social Impact Categories: People + Society

- Concentration of Autho
  - Militarization, Surveilla
  - Imposing Norms and V
- Labor and Creativity
  - Intellectual Property a
  - 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

SHARE

Quick questions break

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

Actionability and 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.”*

Actionability and Transparancy

- 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.*

Actionability and Interpretability

- Interpretability as a field seeks to examine the process of arriving at a particular output

Actionability and Measurement Validity

- Consequential Validity: I.e., “identifying and evaluating the consequences of using the measurements obtained from a measurement model”
  - *Jacobs and Wallach. 2021. Measurement and Fairness*
- Predictive Validity: “the extent to which measurements obtained from a measurement model are predictive of measurements of any relevant observable properties… thought to be related to the construct purported to be measured”
  - Ibid.
- Hypothesis validity: “the extent to which the measurements obtained from a measurement model support substantively interesting hypotheses about the construct purported to be measured”
  - Ibid.

Literature Review

- We search for papers that mention “fair,” “bias,” or “stereotyp*” and which co-occur with either “eval*” or “metric.”
  - Remove irrelevant papers
- Do a literature review of 146 papers from the ACL anthology

Motivation	**R**$_Y$	**R**$_N$
Lack of reliability of existing measures	8	11
Measuring a missing or new bias	8	6
Measuring in a new setting or modality	14	16
Adjusting existing measures[11]	10	10
Measuring in a new language	12	15
No or unclear motivation	7	26
Total	59	84

Table 1: **Motivations provided for new measures.** Absolute counts in our collection (n=146) split into whether the authors discuss reliability (R_Y) or not (R_N).

Question Time