Diagnosing NLP

Sources of Social Harms of NLP

24/July/2024

Zeerak Talat
z@zeerak.org | @zeeraktalat
www: zeerak.org

Past

- B.Sc. Computer Science & M.Sc. in IT and Cognition @ University of Copenhagen
- Ph.D. Computer Science @ University of Sheffield
- Postdoc @ Digital Democracies Institute

Current

- Research Fellow @ Mohamed Bin Zayed University of Artificial Intelligence
- Visiting Researcher @ Alexander von Humboldt Institute For Internet and Society

Future

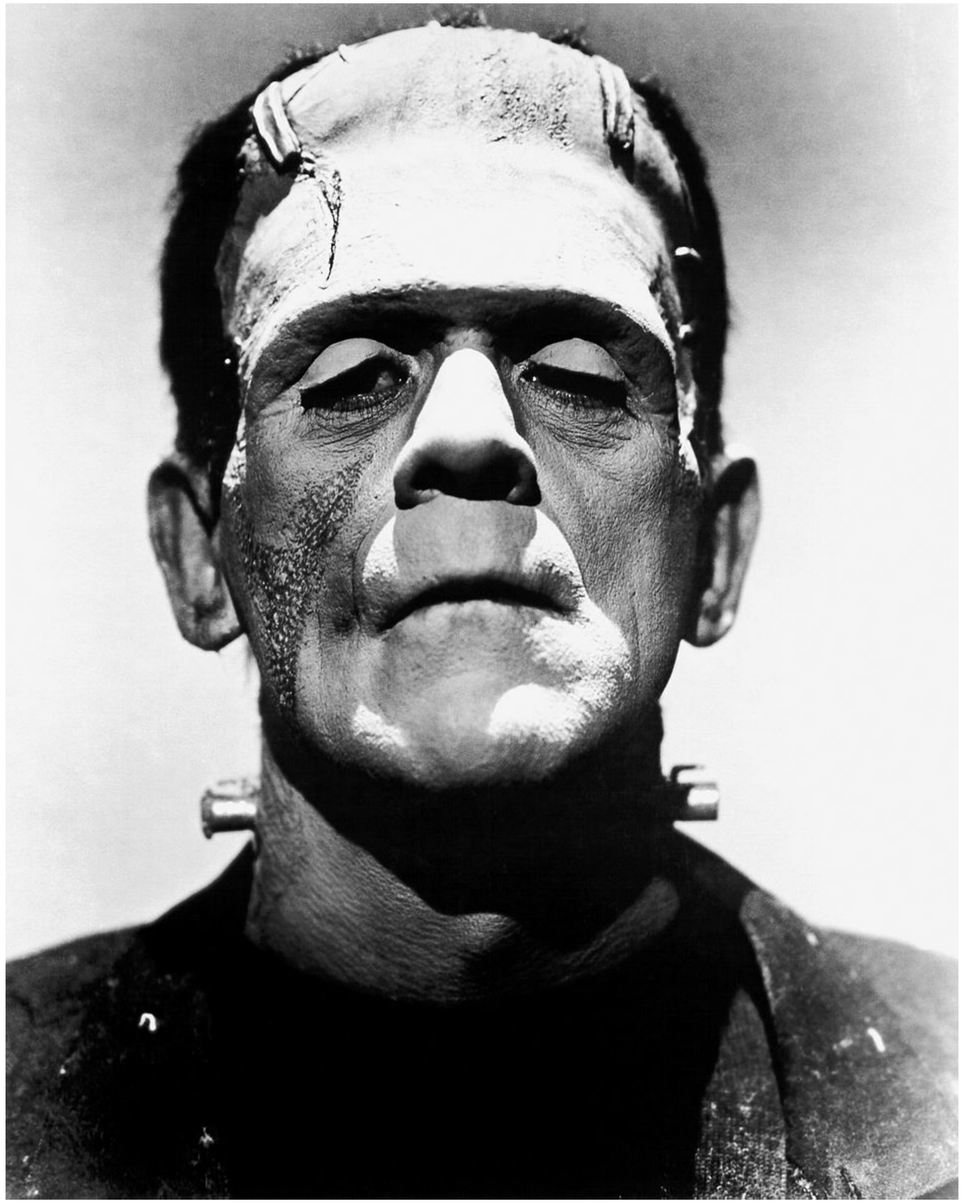
- Chancellor's Fellow (~Assistant Professor) @ the Centre for Technomoral Futures & Edinburgh Informatics - Edinburgh University

The Problem with NLP

What is language?

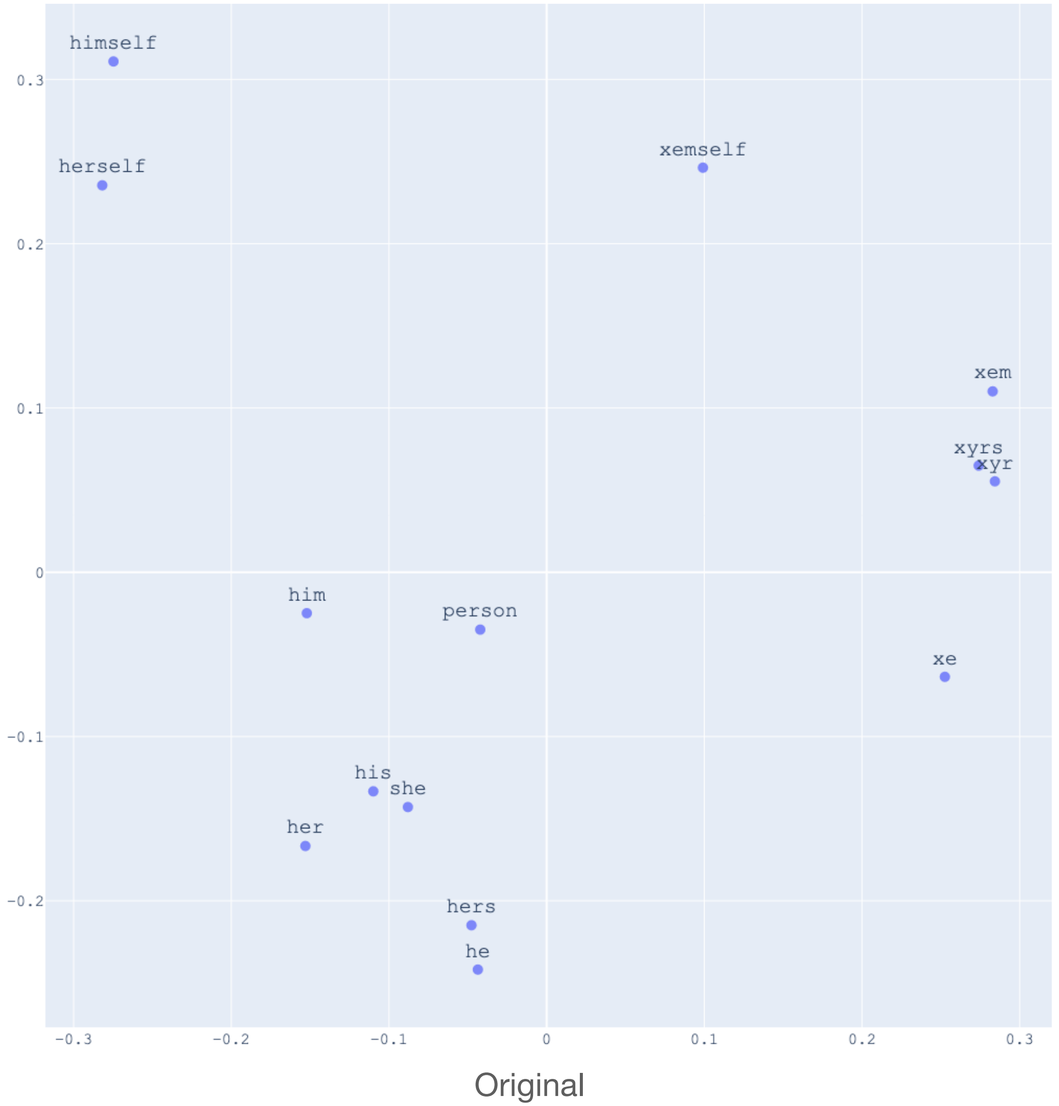
What is language?

How do we use language?

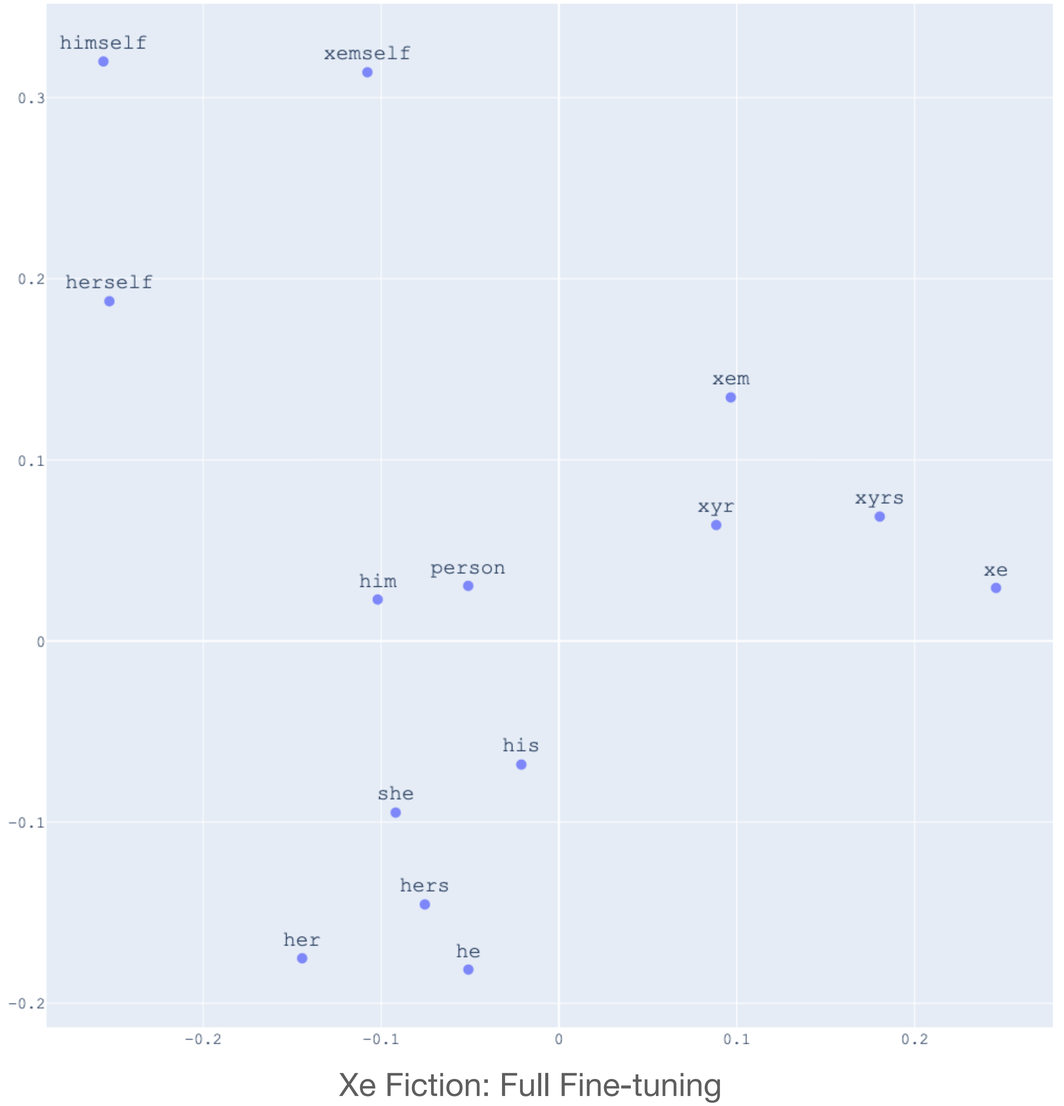


Boris Karloff as Frankenstein’s Monster.
Source: Frankenstein (1931)

Diagnosing NLP: **Fictions**



Original



Xe Fiction: Full Fine-tuning

Talat and Lauscher. 2022. “Back to the Future: On Potential Histories in NLP.”

Broken Language Technologies

Why care about a politics for Natural Language Processing?

Why care about a politics for Natural Language Processing?

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

A Survey on Gender Bias in Natural Language Processing

Gender Bias in Coreference Resolution:
Evaluation and Debiasing Methods

Stereotype and Skew: Quantifying Gender Bias
in Pre-trained and Fine-tuned Language Models

A Survey on Gender Bias in Natural Language Processing

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints

Measuring Bias in Contextualized Word Representations

Why care about a politics f

Mitigating Gender Bias in Natural Language Processing:
Literature Review

Multi-Dimensional Gender Bias Classification

ammer as Woman is to Homemaker?

Identifying and Reducing Gender Bias in Word-Level Language Models

Assessing Gender Bias in Machine Translation – A Case
Study with Google Translate

ler Bias in Abusive Language Detection

Lipstick on a Pig:
Debiasing Methods Cover up Systematic Gender
in Word Embeddings But do not Remove Them

Mitigating Gender Bias Amplification in Distribution by
Posterior Regularization

Examining Gender Bias in Languages with Grammatical Gender

Search titles and abstracts for mentions of terms related to race:

Terms: “race”, “racial”, “racism”

165 potential papers identified

86 Papers excluded for relevance

As they only deal with “racism” as a form of hate speech

or mention “race” as a motivation, related work, or future work

79 papers included

	Collect Corpus	Analyze Corpus	Develop Model	Detect Bias	Debias	Survey/Position	Total
Abusive Language	6	4	2	5	2	2	21
Social Science/Social Media	2	10	6	1	-	1	20
Text Representations (LMs, embeddings)	-	2	-	9	2	-	13
Text Generation (dialogue, image captions, story gen.)	-	-	1	5	1	1	8
Sector-specific NLP applications (edu., law, health)	1	2	-	-	1	3	7
Ethics/Task-independent Bias	1	-	1	1	1	2	6
Core NLP Applications (parsing, NLI, IE)	1	-	1	1	1	-	4
Total	11	18	11	22	8	9	79

Search titles and abstracts for mentions of terms related to socio-economic strata:

Terms: 'social class', 'caste', and 'socio-economic', 'income', 'education', 'occupation', 'white/blue collar', 'upper/middle/lower class'

78 potential papers identified

57 papers excluded for relevance

21 papers included

	Measurement	**Granularity**
Lampos et al. (2014)	Unemployment	Country
Preoţiuc-Pietro et al. (2015)	Occupation	Individual
Flekova et al. (2016)	Income	Individial
Hasanuzzaman et al. (2017)	Income	Individual
Giorgi et al. (2018)	Income, Education	County (census data)
Zamani et al. (2018)	Income, Education, Unemployment	Country-wise
Degaetano-Ortlieb (2018)	Class (high, low)	Individual
Van et al. (2019)	Income, poverty education	State-level (census)
Jawahar and Seddah (2019)	Income, geolocation	Neighbourhood-level
Basile et al. (2019)	Restaurant price	Individual
Ghazouani et al. (2019)	socio-economic status	Mixed
Abraham et al. (2020)	Income, area	Group
Tafreshi et al. (2021)	Income, education	Individual
Abbasi et al. (2021)	Income, education	Individual
Strømberg-Derczynski et al. (2021)	SES (high, mix, unknown)	Aggregated by dataset
van Boven et al. (2022)	Low-income countries	Country
Ngao et al. (2022)	Low-income countries	Country
Grützner-Zahn and Rehm (2022)	GDP	Country
Cole (2022)	Class (high, low)	Individual
Malik et al. (2022)	Caste, occupation	General (bias)
Hržica et al. (2022)	Class (middle)	Group

	Organization	Author Location	Language	Parameters	Model Access	Bias Eval
MT-NLG	Microsoft, NVIDIA	USA	English	530 B	Closed	Smith et al. (2022)
Gopher	DeepMind	USA	English	280 B	Closed	Weidinger et al. (2021b)
ERNIE 3.0	Baidu	China	English, Chinese	260 B	Closed	—
Yuan 1.0	Inspur AI	China	Chinese	245 B	Closed	—
HyperCLOVA	NAVER	Korea	Korean	204 B	Closed	—
PanGu-α	Huawei	China	Chinese	200 B	Closed	—
Jurassic-1	AI21 Labs	Israel	English	178 B	Commercial	—
GPT-3	OpenAI	USA	English	175 B	Commercial	Brown et al. (2020)
LaMDA	Google	USA	English	137 B	Closed	Thoppilan et al. (2022)
Anthropic LM	Anthropic	USA	English	52 B	Closed	Askell et al. (2021)
GPT-NeoX-20B	EleutherAI	Multinational	English	20 B	Open	(Gao et al., 2020; Biderman et al., 2022)
Turing NLG	Microsoft	USA	English	17 B	Closed	—
FairSeq Dense	Meta AI	Multinational	English	13 B	Open	—
mT5	Google	USA	Multilingual	13 B	Open	—
ByT5	Google	USA	English	13 B	Open	—
T5	Google	USA	English	11 B	Open	—
CPM 2.1	Tsinghua University	China	Chinese	11 B	Open	—
Megatron 11B	NVIDIA	USA	English	11 B	Open	—
WuDao-GLM-XXL	Beijing Academy of AI	China	Chinese	10 B	Open	—
WuDao-GLM-XXL	Beijing Academy of AI	China	English	10 B	Open	—
BlenderBot	Meta AI	USA	English	9 B	Open	—
Megatron-LM	NVIDIA	USA	English	8 B	Closed	—
XGLM	Meta AI	Multinational	Multilingual	7 B	Open	—
GPT-J-6B	EleutherAI	Multinational	English	6 B	Open	(Gao et al., 2020; Biderman et al., 2022)

Talat, Z. et al. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models*, 26–41.

A Survey on Gender Bias in Natural Language Processing

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models

A Survey on Gender Bias in Natural Language Processing

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

Measuring Bias in Contextualized Word Representations

Mitigating Gender Bias in Natural Language Processing: Literature Review

Multi-Dimensional Gender Bias Classification

Man is to Computer Programmer as Woman is to Homemaker?

Identifying and Reducing Gender Bias in Word-Level Language Models

Assessing Gender Bias in Machine Translation – A Case Study with Google Translate

Gender Bias in Abusive Language Detection

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender in Word Embeddings But do not Remove Them

Mitigating Gender Bias Amplification in Distribution by Posterior Regularization

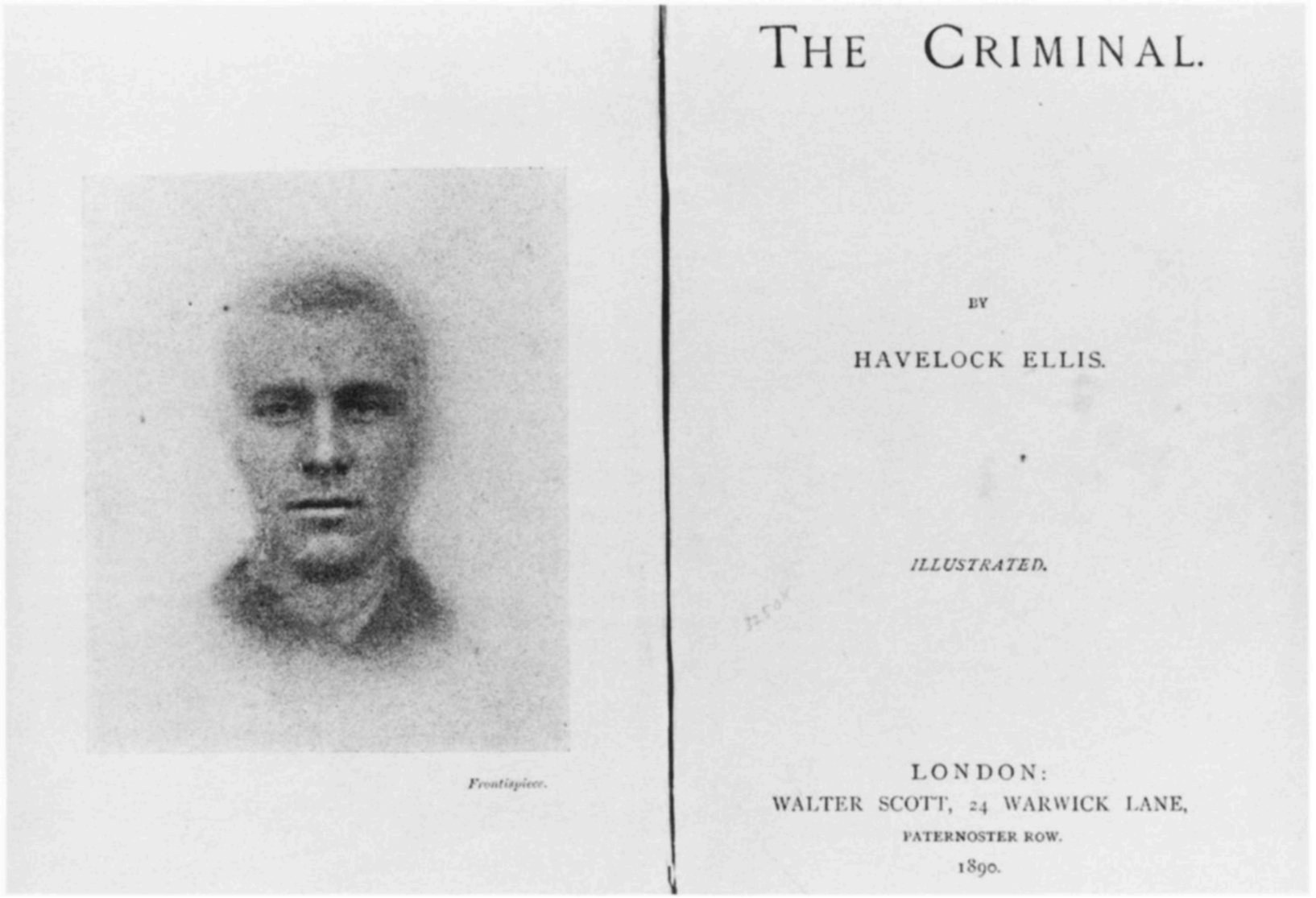
Examining Gender Bias in Languages with Grammatical Gender

"[L]anguage models' attitudes about AAE are even more negative than the most negative experimentally recorded human attitudes about African Americans, i.e., the ones from the 1930s."

Hoffman et al., Dialect prejudice predicts AI decisions about people's character, employability, and criminality. ArXiv. 2024.

Deployment and use of NLP

Source: WWW's "historical" logo, created by Robert Cailliau in 1990. Wikimedia.



Frontispiece.

THE CRIMINAL.

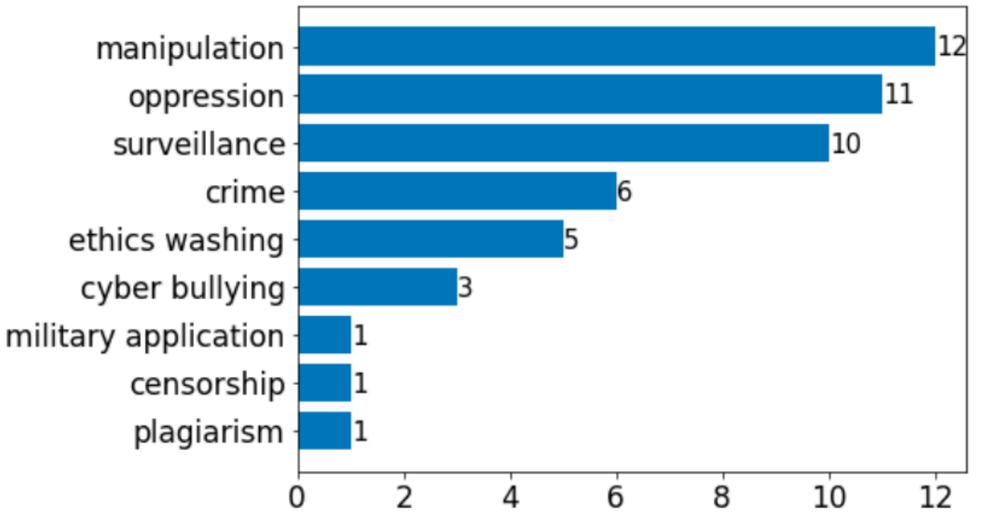
BY

HAVELOCK ELLIS.

ILLUSTRATED.

LONDON:
WALTER SCOTT, 24 WARWICK LANE,
PATERNOSTER ROW.
1890.

A Galtonoian Composite
Source: Alan Sekula (1986). The Body and the Archive. October. MIT Press



Distribution of codes for harms identified by participants of the tasks in NLP they work on.

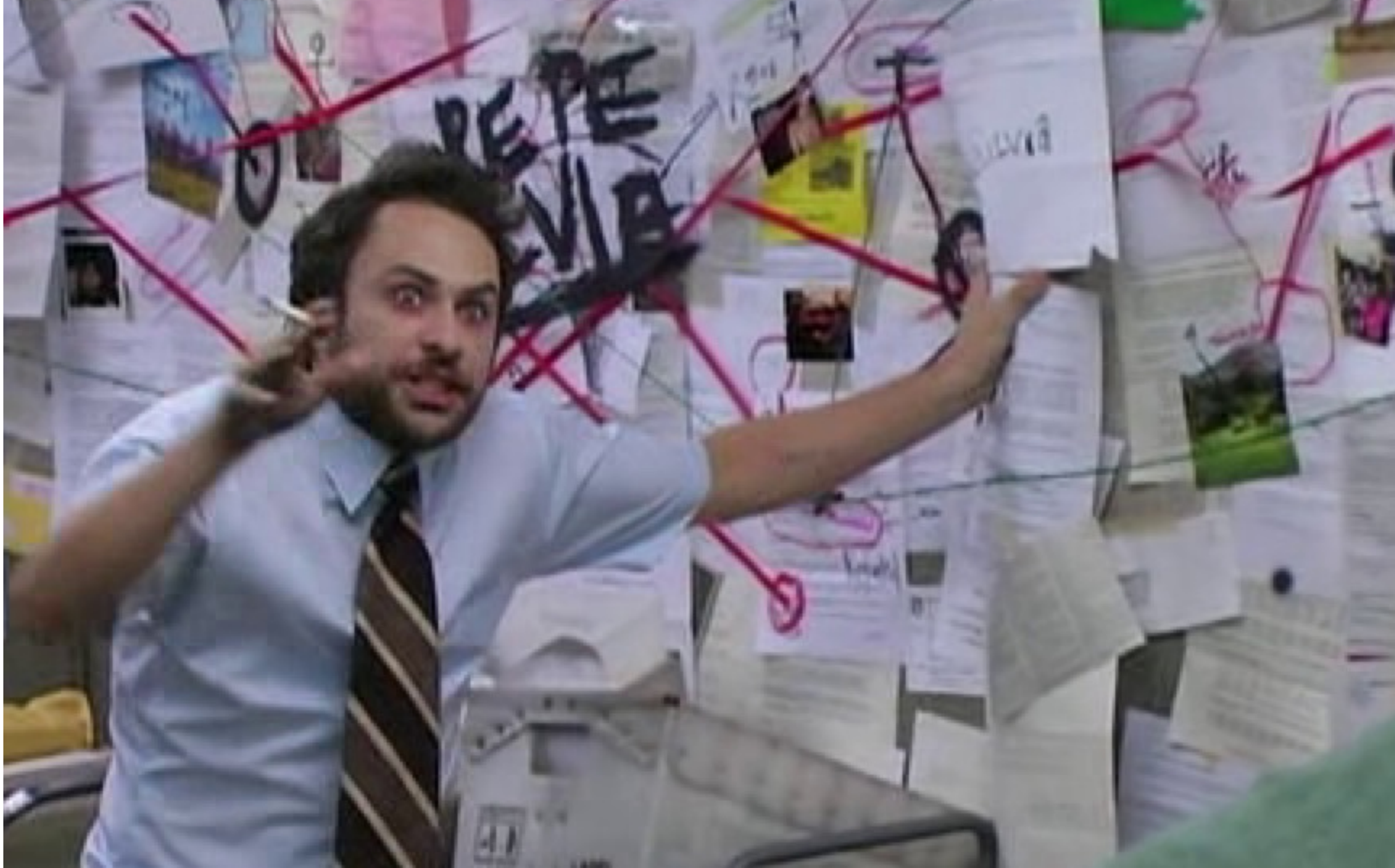
Area	vulnerability	harms
NLP Applications	4.3	crime, oppression, manipulation
Ethics and NLP	4.1	ethics washing, surveillance, plagiarism, oppression
Psycholinguistics	4.0	cyber bullying, oppression
Generation	3.8	crime, cyber bullying, manipulation, oppression, ethics washing
Dialogue Systems	3.8	surveillance, crime
MT and Multilinguality	3.3	surveillance, crime
ML for NLP	3.2	military application, manipulation, oppression
Resources and Evaluation	3.1	ethics washing, surveillance, manipulation
Interpretability	3.0	ethics washing, manipulation
Information Extraction	2.8	surveillance, censorship
IR and Text Mining	2.7	surveillance

Average score for vulnerability across ACL areas (with at least three answers) the participants work on and their associated harms.

Kaffee et al., 2023. “Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing.”
In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics.

Ideas about separating, purifying, demarcating and punishing transgressions have as their main function to impose system on an inherently untidy experience.

"Purity and Danger. An Analysis of the Concepts of Pollution and Taboo." Mary Douglas (1978)



Concluding Remarks

- Privacy
- Inclusion (in the Diversity and inclusion sense)
- Individual rights

- Birhane, Abeba, and Zeerak Talat. 2023. “It’s Incomprehensible: On Machine Learning and Decoloniality.” Pp. 128–40 in *Handbook of Critical Studies of Artificial Intelligence*, edited by S. Lindgren. Edward Elgar Publishing.
- Curry, Amanda Cercas, Zeerak Talat, and Dirk Hovy. 2024. “Impoverished Language Technology: The Lack of (Social) Class in NLP.” in Proceedings of LREC-COLING 2024.
- Douglas, Mary. 1978. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Repr. London: Routledge.
- Field, Anjalie, Su Lin Blodgett, Zeerak Talat, and Yulia Tsvetkov. 2021. “A Survey of Race, Racism, and Anti-Racism in NLP.” Pp. 1905–25 in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics.
- Haraway, Donna. 1988. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” *Feminist Studies* 14(3):575–99. doi: 10.2307/3178066.
- Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. “Dialect Prejudice Predicts AI Decisions about People’s Character, Employability, and Criminality.”
- Kaffee, Lucie-Aimée, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. “Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing.” Pp. 13977–98 in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics.
- Sekula, Allan. 1986. “The Body and the Archive.” *October* 39:3. doi: 10.2307/778312.
- Talat, Zeerak, and Anne Lauscher. 2022. “Back to the Future: On Potential Histories in NLP.”
- Talat, Zeerak, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. ““Disembodied Machine Learning: On the Illusion of Objectivity in NLP.””
- Talat, Zeerak, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. “You Reap What You Sow: On the Challenges of Bias Evaluation under Multilingual Settings.” Pp. 26–41 in *Proceedings of BigScience episode #5 – workshop on challenges & perspectives in creating large language models*. virtual+Dublin: Association for Computational Linguistics.