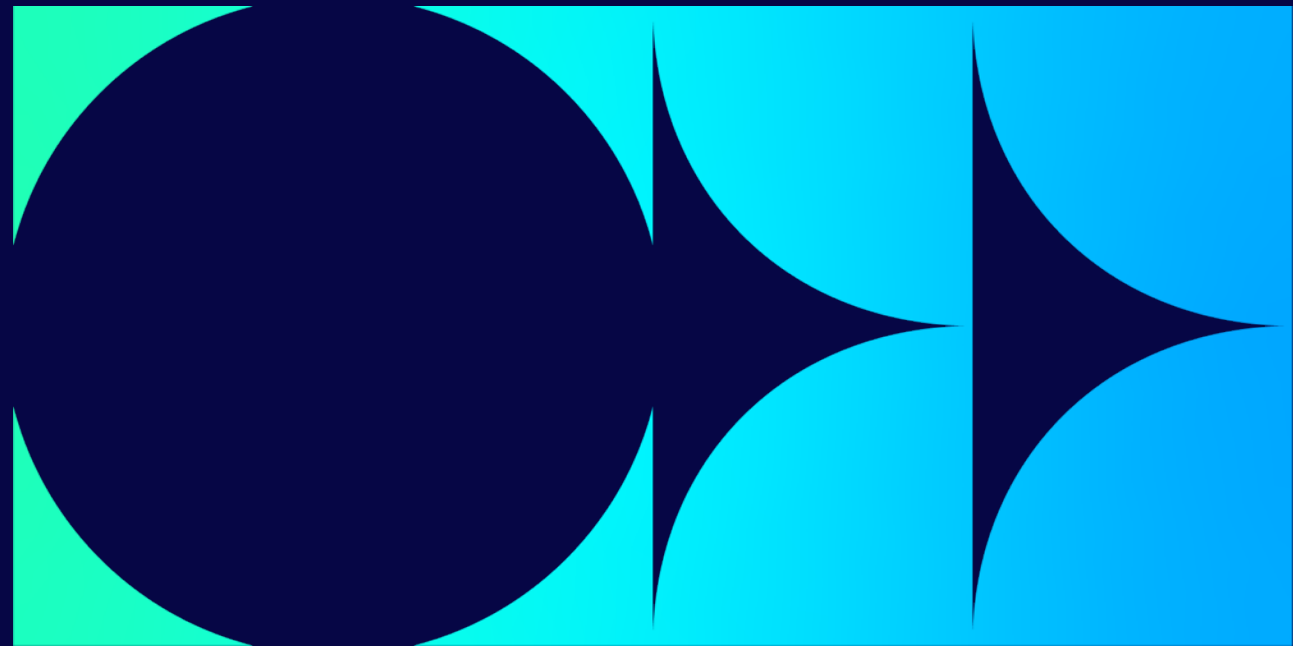
A Multidisciplinary Lens of Bias in Hate Speech

Paula Reyero Lobo

Knowledge Media Institute
The Open University, UK

7 November 2023

Paula Reyero Lobo, Joseph Kwarteng, Mayra Russo, Miriam Fahimi, Kristen Scott, Antonio Ferrara, Indira Sen, Miriam Fernandez



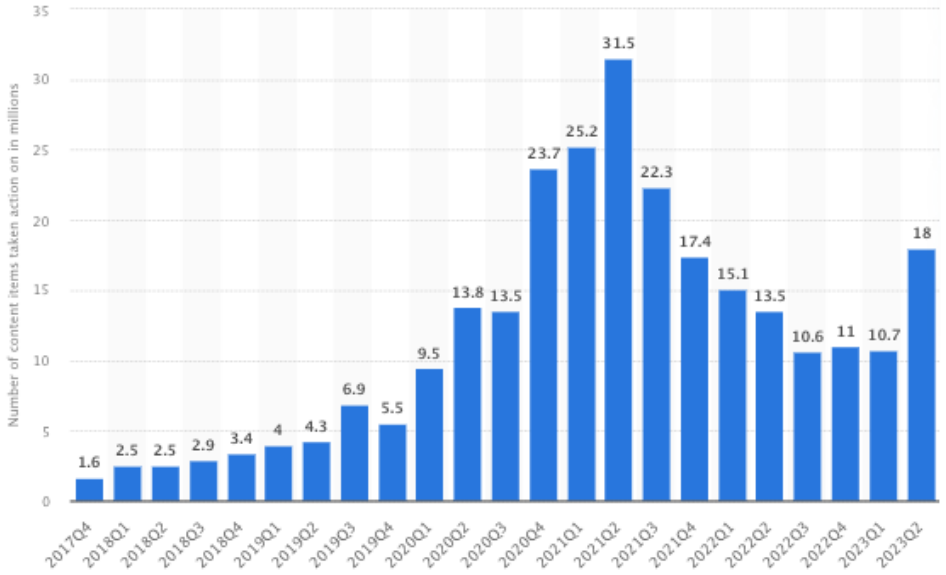
The Open University

Funded by the European Union

Introduction

Why do we need automated tools in content moderation?

18 Million pieces of hate speech removed on Facebook between April and June 2023



Statista research Department, 20 September 2023

Amnesty International UK

Press releases

TWITTER: Anti-LGBTQ+ hate speech surging on platform under Elon Musk

09 Feb 2023, 07:11pm

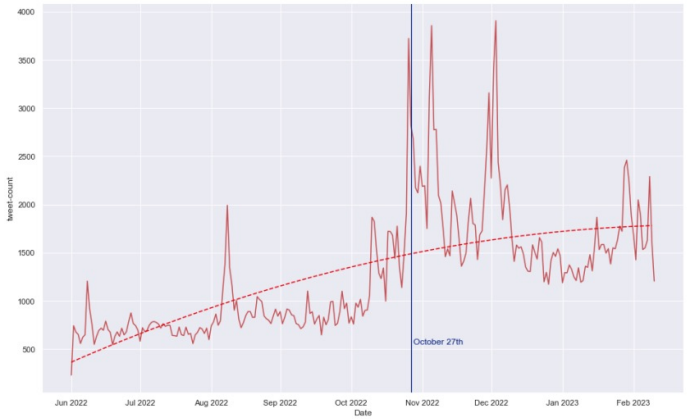


Figure 1: volume of potentially antisemitic Tweets over time, June 2022 – February 2023

Institute of Strategic Dialogue

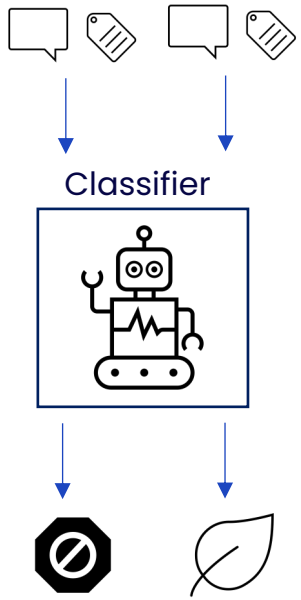
Growing prevalence of hate speech in X in 2023

The Open University

Introduction

Automated detection systems vary their score depending on the specific demographic characteristics of the speaker or target of hate.

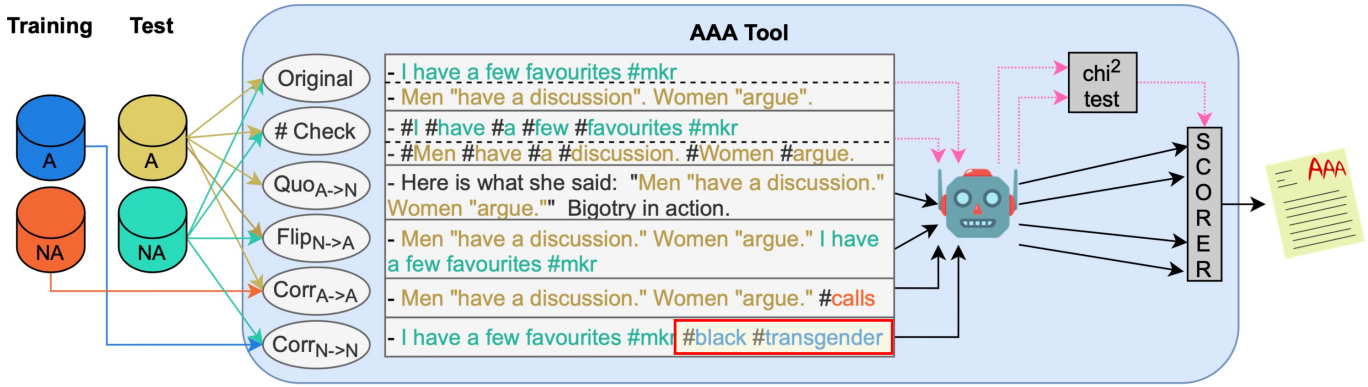
Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, Yejin Choi. 'Challenges in Automated Debiasing for Toxic Language Detection'. ACL 2021.



The challenge is to optimize for performance and **bias reduction**!

Related Work

The interrelated issues of fairness, model robustness, and disproportionate harms are often addressed in isolation.



Agostina Calabrese, Michele Bevilacqua, Björn Ross, and Roberto Navigli. AAA: Fair Evaluation for Abuse Detection Systems Wanted. WebSci 2021.

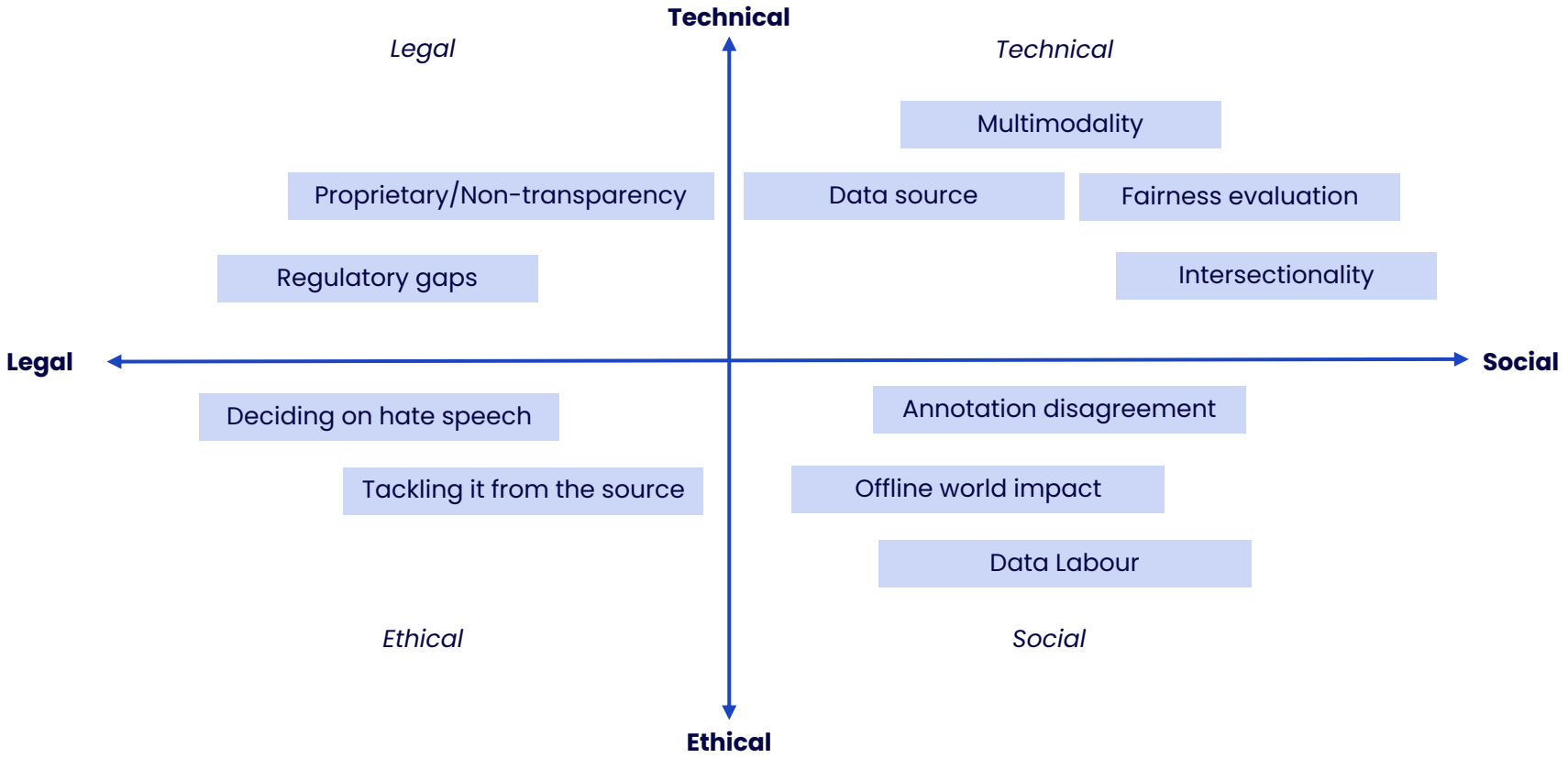
	Functionality	Example Test Case	Gold Label	n
Negation	**F14**: Hate expressed using negated positive statement	"[IDENTITY] lives do not matter."	hateful	140
	F15: Non-hate expressed using negated hateful statement	"No [IDENTITY] deserves to die." "We shouldn't let [IDENTITY] suffer."	non-hate	133
Phrasing	**F16**: Hate phrased as a question	"Why is it that all [IDENTITY] are repulsive?"	hateful	140
	F17: Hate phrased as an opinion	"I think that [IDENTITY] should all be killed."	hateful	133
Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	126
	F19: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	189
Counter speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	173
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	141

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, Jane B. Pierrehumbert. HateCheck: Functional Tests for Hate Speech Detection Models. ACL 2021.

The real impacts of bias on people and society are difficult to understand with this measures alone.

Our approach

We design a data challenge to involve interdisciplinary participation to discuss existing and underexplored issues.



We identify critical bias *facets* that must be considered when deploying hate speech detection in society.

Technical

Intersectionality

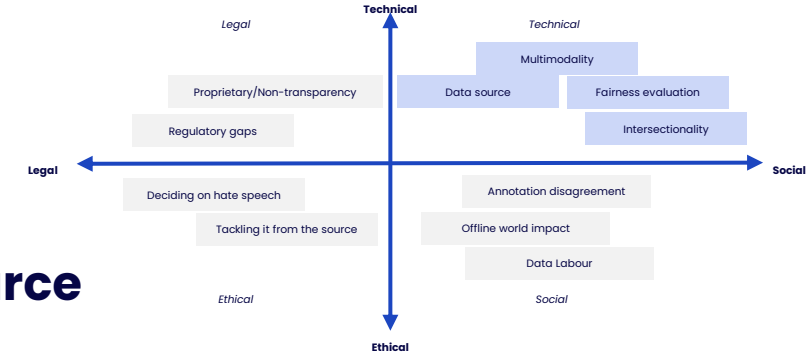
Different social identities overlap, creating unique systems of discrimination.

Make detection tools *intersectionally-sensitive*: diverse and representative data, feature engineering considering intersectional characteristics, evaluate disproportionate harms towards any marginalized group.

Fairness evaluation

Bias investigations address unintended identity bias, where non-harmful content containing identity terms is misclassified as hateful.

Include more nuanced and interpretable fairness metrics that are grounded on real-world hams and deeper investigations into the source of bias.



Data source

Hate speech detection rely on datasets mainly elaborated from large volumes of social data.

Valued-based data collection and ethical data annotation practices could already undercut quality issues down the pipeline.

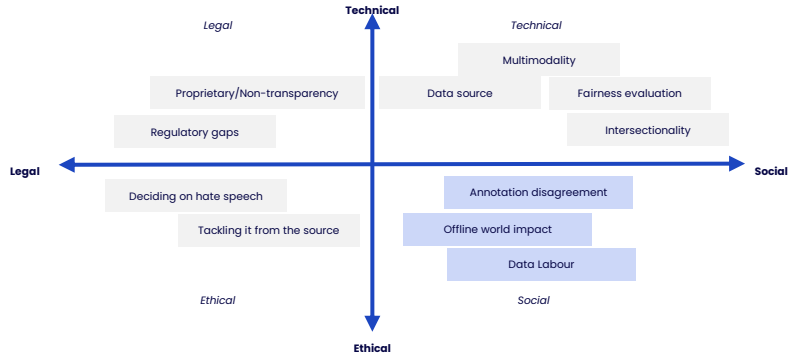
Multimodality

Despite the strong focus on textual data, the multimodal nature of the problems poses additional challenges.

Human-centered evaluation is essential to respond to the preferences of a diverse sample of users interacting with these systems.

The Open University

Social



Annotation disagreement

Annotating hate speech involves subjective judgments and depends on personal experiences.

Retention of annotator-level documentation to uncover hidden patterns or nuances in data that would be missed.

Offline world impact

Hateful content has led to violence and affects psychological well-being, but excessive moderation can veer into censorship.

The amount of moderation or lack thereof requires careful design choices and thinking about trade-offs.

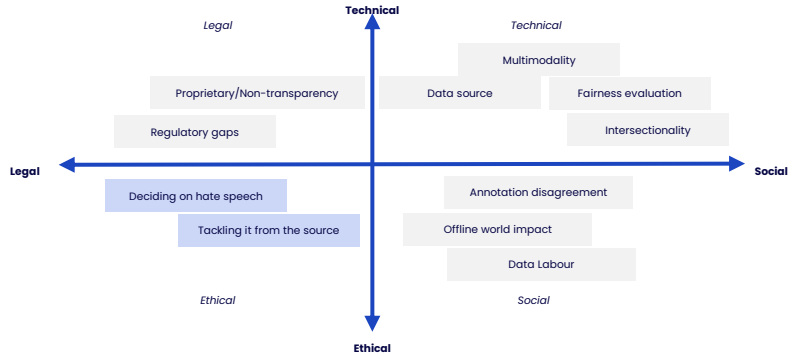
Data Labour

Hate speech detection requires data labor for the creation of data collection, aggregation, labelling, content disambiguation, and content moderation.

These frameworks should empower laborer by enabling a transparent communication with their contractors, feedback channels, sustained mental counsel, and better remuneration.

The Open University

Ethical

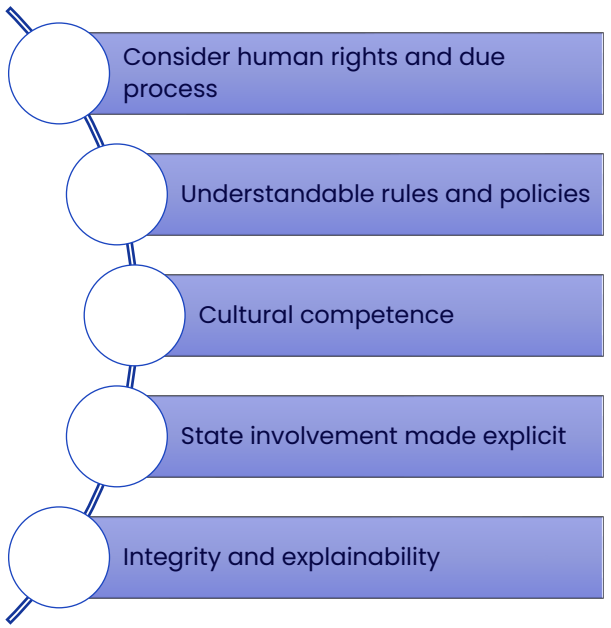


Deciding on hate speech

What is considered hateful varies across cultures, location, social groups, and contextual factors such as time of the day.

Ensuring responsible platform moderation requires to take a human rights approach.

Santa Clara Principles



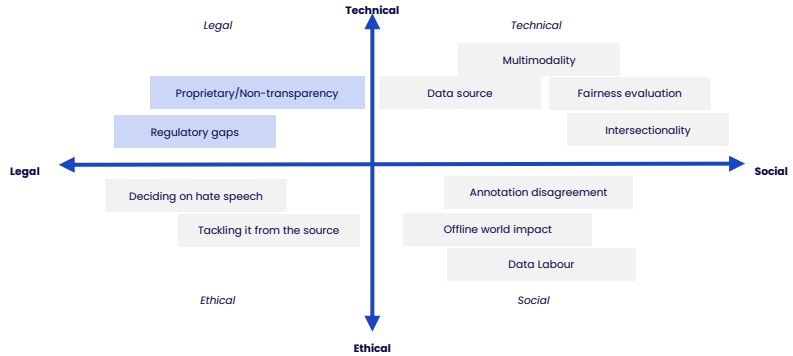
Tackling it from the source

Hate speech is a co-constructive process between an actor and receiver, entangled in their positionalities and social embedding.

Other disciplines such as those from gender studies can help dismantle how structural forms of discrimination enter collective and individual speech.

The Open University

Legal



Proprietary and Non-transparency

Proprietary rules and non-transparent mechanism to moderate content introduce a corporate bias.

Community-driven interventions and guidelines give the possibility to shape and intervene in how our content is moderated.

Regulatory gaps

There is a prevailing legal understanding of discrimination to categorize hate speech.

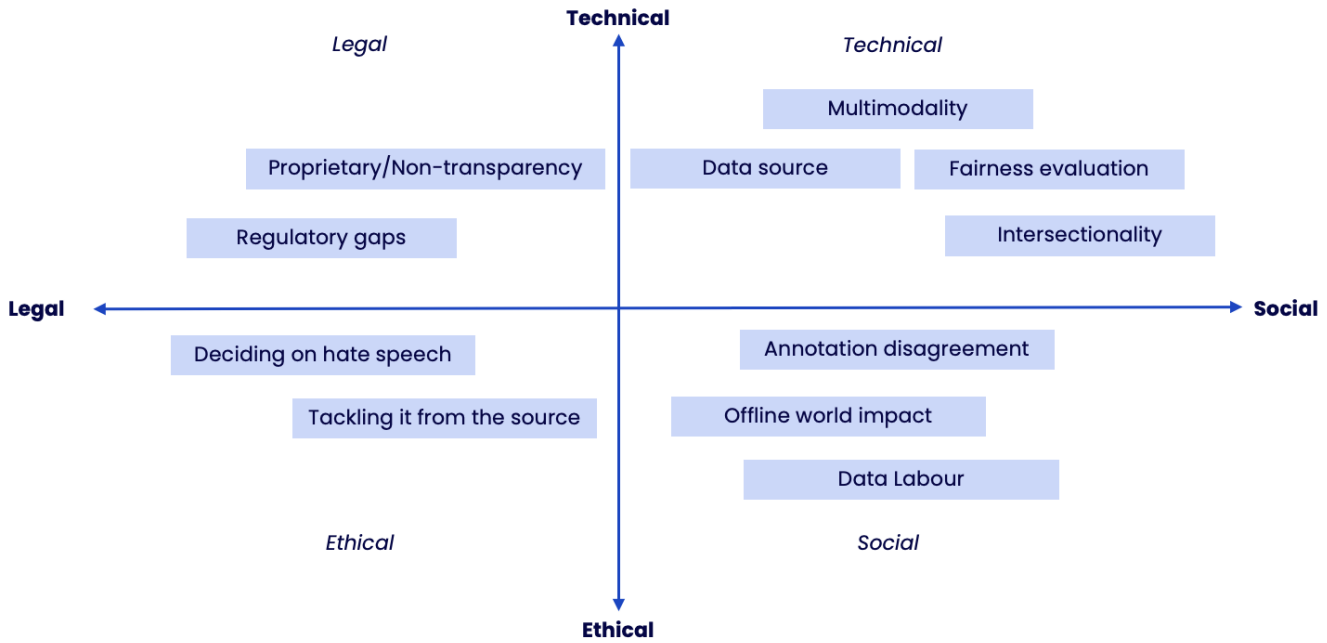
Other forms of discrimination such as *class*, *speciesist* and *decolonial* biases pose gaps in the current regulation.

The Open University

Conclusion

A nuanced understanding of bias is key for ensuring

- Technical guarantees
- But also the interrelated ethical,
- Societal,
- And legal aspects of the problems surrounding hate speech detection



Bias facets at the multidisciplinary level. Each facet shifts along technical, social, ethical, and legal axes and cannot be treated in isolation.

The Open University

Tackling bias in hate speech detection requires multidisciplinary methods

A Multidisciplinary Lens of Bias in Hate Speech
Reyero Lobo, P., Kwarteng, J., Russo, M., Fahimi, M., Scott, K., Ferrara, A., Sen, I. & Fernandez, M.

Contact

paula.reyero-lobo@open.ac.uk

www.linkedin.com/in/paula-reyero-lobo-116449170

@paulareyero1

The Open University

Funded by the European Union