Oh, the humanity! A human-centric approach to social bias research in Natural Language Processing

Eddie L. Ungless

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2025

Abstract

Much current research into social bias in Natural Language Processing (NLP) – that is, the tendency for NLP technologies to reflect human biases such as sexism and homophobia in the relative probability of different outputs – suffers from relying on a superficial understanding of the problem. The issue of social bias is treated as a mathematical kink that needs to be ironed out, after which the harm that the model does will be irrefutably reduced – a form of algorithmic idealism. Social bias is seen as an unfortunate result of "dirty data" and "data imbalance", and practitioners typically focus on addressing social bias through changes at training or inference to counteract these data issues. Little regard is given to the human aspect: to the myriad normative choices made by those who develop these systems; the beliefs of those who deploy them; nor to the response of those impacted by these technologies, all of which will influence how social bias is actually experienced. Operationalising bias as a quantifiable metric allows for at-scale evaluation that keeps pace with the rapid development of new NLP technologies. However, I argue this superficial understanding of social bias will lead us to superficial and ultimately ineffective solutions, which ignore the role of human behaviour in determining the harm done by technology. As I demonstrate, heuristic attempts at social bias mitigation often end up doing more harm than good.

In my thesis, I advocate for a human-centric approach to measuring and mitigating social bias in NLP, one which focuses on human choices, human identities and human behaviour, to give a more complete understanding of the true impact of NLP technologies. A human-centric approach treats social bias as a socio-technical problem, and casts its net widely over a broad range of stakeholders, sources of bias, and demographic attributes. My proposed approach is underpinned by five maxims: see technology as part of a socio-technical system; consider many sources of bias; focus on the impact on people and how they respond; be driven by social science theory and community knowledge; address a broad range of demographics.

I present my work as four case studies across three tasks which demonstrate the benefits of this approach. I consider harms against marginalised (primarily queer) identities through social bias in sentiment analysis tools, text-to-image (TTI) models and social media recommender and moderation algorithms (namely on TikTok), finding that heuristic attempts to reduce social bias often do more harm than good, and that the public form complex beliefs about NLP technologies. In all my work, I address social biases in publicly available or public facing tools, as these typically have a broader impact than state of the art models. I focus primarily on harms done

to the LGBTQ+ community, in part because of personal relevance, but also because it provides an opportunity to demonstrate the benefits of an approach that considers demographic qualities beyond a binary. There are no binaries in nature – in human identity – yet much social bias research hinges on treating demographics as such. In doing so I contribute significantly to our understanding of queerphobia in NLP.

Ultimately I argue – despite the title of this thesis – that the best approach to social bias research is one that switches focus from social bias to real-world harms. Social bias has been used as a proxy for harm, but as I argue, it is often a very poor one. Changing our focus to harms necessitates defining specific use contexts (real or imagined). The "meaning" of a difference in probability will be context dependent, as will how this difference is interpreted by those impacted by the model. I am far from the first to critique current practices, and the disconnect between social bias and harm; I amplify the message, and enrich it with five clear maxims that improve the validity of social bias research. To leave the human context out of the equation when measuring harm is nonsensical, yet for too long the field of NLP has attempted to do exactly that. *Oh, the humanity!*

Acknowledgements

A sincere thank you to my family, friends, lovers and supervisors. In particular Björn Ross, your support has been invaluable. Thank you also to my examiners Alex Taylor and Maarten Sap for an enjoyable viva and important critique.

It is also my sad duty to acknowledge that The University of Edinburgh has failed to uphold standards of academic integrity with regards to research and media related to trans people, and I wish to distance my own work from that which has been inappropriately given a platform at this university.[1] It is my firm belief that The University of Edinburgh is institutionally transphobic, as evidenced by it allowing the screening of a sensationalised "documentary" with no academic merit in the name of academic freedom, and the appointment of a trans-exclusionary activist to the position of Rector. The University has also set the absurd requirement that events related to trans issues be presided over by a "neutral chair", as if one can be neutral in the face of oppression.

Attempts to critique these decisions are silenced by the university.[2] The UCU branch co-presidents seemed to be required to issue an absurd apology for calling out transphobia - it is *never* for the oppressor to define what counts as oppression. Freedom of expression should not mean freedom from criticism, and using "academic freedom" to justify silencing trans staff, students and their allies *is* an example of transphobia. Historic parallels with the suppression of discussion of other queer identities make this clear. Students who stand up to the "gender critical" movement are given no support by the university and are instead thrown to the wolves. I expect the same to happen to me in writing this statement, but I refuse to be silent.

I do not believe The University of Edinburgh is a safe place for trans people to study, within the context of relative (but fragile) safety within the UK,[3] and I discourage anyone from following me in pursuing a PhD at this institution until reform occurs.

Whilst I have had the unwavering support of all staff with whom I have worked directly, in the end I have completed my thesis not because of the environment at The University of Edinburgh, but in spite of it.

---

[1] https://blogs.ed.ac.uk/staffpridenetwork/2023/04/14/is-screening-adult-human-female-an-opportunity-for-respectful-debate-and-discussion/

[2] https://blogs.ed.ac.uk/staffpridenetwork/2024/02/14/spn-committee-statement-regarding-the-appointment-of-simon-fanshawe-to-rector-of-the-university-of-edinburgh/

[3] cf. https://www.stonewall.org.uk/about-us/news/new-data-rise-hate-crime-against-lgbtq-people-continues-stonewall-slams-uk-gov-

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

The structure and content of my background section is similar to a whitepaper on ethical research with large language models (LLMs) on which I am first author, namely Ungless et al. (2024) (indeed, the structure of that paper was inspired by my decision to structure my thesis background section around stages in a model life cycle). This was also published as a "pocket guide" as Ungless et al. (2025d). Work on these papers was done in collaboration with other authors, but sections on which I was the primary author will likely have some overlap to this thesis. Specifically, I was the primary author of the sections on best practice when working with stakeholders; issues of consent and safety when compiling data; cleaning data; choice of target labels; fairness and debiasing techniques; harm evaluation; and interventions at inference, and these overlap in content with Sections 2.1, 2.2 and Chapter 3 in this thesis. At no point do I take insights from my co-authors and include them in this thesis as my own.

In places, I lift insights and writing from my blog `mxeddie.github.io`. Specifically, when discussing how alt-text reflects the observer's beliefs about a community, which will then be learned by the model, and the use of non-recommended language in this data in Section 2.1.2.

In places, I have lifted text from talk descriptions I have written. Namely, my conclusion features text from my Controversies in the Data Society[4] talk summary. My abstract features text from my Social Data Science Hub[5] talk summary.

In places, I make reference to arguments from other papers on which I have been joint first author, namely Sigurgeirsson and Ungless (2024); Bird et al. (2023); Goldfarb-Tarrant et al. (2023). I phrase these references as e.g. "elsewhere I have argued" but I do not wish to take sole credit for these insights - this phrasing is intended to draw attention to the fact I am referencing these papers as a joint first author, distinct from my other citations.

I have included four of my published papers in Chapters 4-7, with minor changes for consistency and clarity. I have included novel introductions in place of the original abstracts, and I have added Learnings sections to explain the direction the papers led

[4] `https://www.wiki.ed.ac.uk/pages/viewpage.action?pageId=618201791`
[5] `https://sds-hub.ed.ac.uk/seminars/eddie-ungless/`

me to follow in my thesis. I use the first person plural in these Chapters both to maintain consistency with my published work and to acknowledge that these sections were the product of collaboration, although I was in all cases first author. My contribution to these papers entailed being primarily responsible for the following tasks:

- Chapter 4: Conceived and designed the analysis; Collected the data; Performed the analysis; Wrote the paper. **Published as:** Eddie Ungless, Björn Ross, and Vaishak Belle. 2023a. Potential Pitfalls with Automatic Sentiment Analysis: The Example of Queerphobic Bias. *Social Science Computer Review.* 41(6), pages 2211-2229

- Chapter 5: Conceived and designed the analysis; Collected the data (survey and interview); Performed the analysis (survey and interview); Wrote the paper. **Published as:** Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023b. Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Pages 7919–7942

- Chapter 6: Conceived and designed the analysis; Collected the data; Performed the analysis; Wrote the paper. **Published as:** Eddie L. Ungless, Nina Markl, and Björn Ross. 2025b. Experiences of Censorship on TikTok Across Marginalised Identities. In *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media, ICWSM 2025*, Copenhagen, Denmark.

- Chapter 7: Conceived and designed the analysis; Collected the data; Performed the analysis; Wrote the paper. **Published as:** Eddie L. Ungless, Nina Markl, and Björn Ross. 2025c. Le$bean or lesbian? marginalised users' motivations for obfuscation on TikTok. *Behaviour & Information Technology.*

Eddie L. Ungless

Lay Summary

In this thesis, I argue that current research on social bias in Natural Language Processing (NLP) (approximately, AI technology that deals with human language) is too simplistic and focused on technical fixes that miss the bigger picture. Social bias refers to how these language technologies can reflect harmful human biases, like sexism or homophobia, by favouring certain outputs over others. However, many proposed solutions fail to consider the real-world impact on those affected by these biases, so are ultimately ineffective.

I argue that effective approaches to measuring and preventing bias must focus on human behaviours, which shape the language technologies and their eventual impact. I propose a human-centred approach to measuring social bias that considers both the technical and social factors. For example, by surveying people from marginalised communities to understand how they interact with a potentially harmful technology, such as a biased social media recommender AI. I focus on harms to LGBTQ+ individuals, both because of personal relevance and because it highlights the limitations of treating social bias in a binary way (e.g., comparing the treatment of "male" to "female", "gay" to "straight") which falls apart when you consider LGBTQ+ identities.

My human-centred approach to measuring social bias has five key rules and I illustrate these rules with four case studies. In my first case study I show how automated sentiment analysis technologies (designed to predict the sentiment or emotional attitude of a piece of text) are biased against LGBTQ+ identity terms, some giving wildly different sentiment scores when these terms are present. I discuss how apparent attempts to reduce bias in two of the technologies, by ignoring certain identity terms, fall short of being effective. In my second case study I show how image creation AI (such as DALL·E) produce harmful images when prompts contain trans and nonbinary identity terms. The images play into stereotypes such as the sexualisation of trans identities. I survey the affected community and find frustration at existing solutions and a desire for greater control. In my third and fourth case studies I look at the experiences of marginalised people, particularly LGBTQ+ users, on TikTok, a platform widely suspected to have very biased recommender and moderation AI (known together as "the algorithm"). In a survey of hundreds of users, I find widespread concerns about unfair censorship. Some users resort to using coded language to avoid this censorship, although this does create an opportunity for creativity.

Ultimately, I argue for a shift in focus: instead of looking at "social bias", we

should look at the real harms caused by these language technologies. “Social bias” is often used as a stand-in for harm, but it does a poor job of this. The real harm depends on the context in which the language technology is used, how people are affected by it and how they respond. Ignoring the human context when evaluating these technologies is limiting and ultimately unhelpful. I am far from the first to criticise current research into social bias; I can only hope that adding my voice to the chorus will amplify the message, and enrich it with five clear rules that improve the quality of social bias research.

Dedicated to the students of Gaza who have been denied the opportunity to learn through death or destruction.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

It has been well established that natural language programming (NLP) technologies suffer from social bias; by the late 2010s, this was being addressed not only in academic work (Bolukbasi et al., 2016; Barocas et al., 2017; Blodgett and O'Connor, 2017; Dixon et al., 2018; Hovy and Spruit, 2016; Kiritchenko and Mohammad, 2018; Mitchell et al., 2019; Rudinger et al., 2018; Sheng et al., 2019; Zhao et al., 2018, *i.a.*) but also public awareness campaigns such as those from Joy Buolumwini and the Algorithmic Justice League,[1] and tech journalism from magazines such as Wired (Knight, 2019) and Vice (Thompson, 2017). Though social bias measurement had become an established subfield within NLP, there were prevalent methodological issues, comprehensively laid out in Blodgett et al.'s blistering indictment of the state of contemporary social bias research in NLP (Blodgett et al., 2020). They highlighted that the papers' motivations were often vague and poorly thought out, and that the operationalisation of bias failed to engage with literature from outside NLP, and did not match these (vaguely defined) motivations.[2]

As Blodgett et al. (*ibid*) identified in their survey, a frequent issue with contemporary work was the failure to integrate relevant social science research, particularly from psychology and linguistics (see also Wallach (2014); Larson (2017); Schnoebelen (2017)). The fact that NLP models are trained on human language data was often treated as incidental by those evaluating bias (Schnoebelen, 2017): very little consideration was given to the complex relationship between identity and language (Blodgett et al., 2020). Thus concepts that are highly relevant to any discussion on language, identity and harm such as indexing and marked identity (Bucholtz and Hall, 2004,

[1] https://www.ajl.org/

[2] Although this paper is very well-known (1370 citations as of March 2025) the issues that Blodgett et al. identified continue to plague the field to this day (Goldfarb-Tarrant et al., 2023).

2005); benevolent stereotypes (Glick and Fiske, 1996); systemic bias (Feagin, 2013; Feagin and Bennefield, 2014) and equity versus equality (Davis et al., 2021; Cook and Hegtvedt, 1983), were typically overlooked. I return to these topics in Section 2.3.

Further, it was often unclear how the particular bias that the practitioners were measuring would actually impact the affected communities (Blodgett et al., 2020; Crawford, 2017). For example, Blodgett et al. (2020) criticise use of differences in sentiment score for text that *mentions* specific social groups to measure bias by authors who state they are motivated to prevent harm caused by differences in score for text *by* different social groups (a line of thinking which Cabello et al. pick up in 2023). Further, machine learning (ML) models were often evaluated independently of use context, which meant bias was often evaluated independently of harm (Selbst et al., 2019). Yet this work was typically motivated by a stated desire to prevent harm (Blodgett et al., 2020).

These methodological issues, related to how bias was conceptualised and operationalised, meant that work on bias measurement in academia and industry often had issues affecting their validity that undermined attempts at mitigation (Gonen and Goldberg, 2019; Davis et al., 2021). For improvements in a metric to be equated with meaningful change and a reduction in harm, it must be clear what the metric is measuring; for example, if neutralising embedding similarity does not actually eradicate the issue of gender bias, this suggests embedding similarity is a poor proxy for this harm (Gonen and Goldberg, 2019).

Addressing these initial concerns – that research should integrate social science theory and be clearly connected to harms – formed the foundation of my own approach to social bias measurement, which was further formalised whilst conducting my research. In the following section I set out my positionality, which informed how I approached this research, and define its scope (including defining key terminology).

1.1 Reflexivity and Positionality Statement

Herein I provide a positionality statement, to communicate how my lived experiences inform the work I present. Whilst such statements are common in social science research and increasingly present in human-computer interaction (HCI) research (Liang, 2021; Liang et al., 2021)(e.g. Karizat et al. (2021); Are (2023); Simpson and Semaan (2021); Haimson et al. (2021) *i.a.*), they are rarely present in NLP research, perhaps because of fear it will impact the (illusion of) objectivity in computer science (Talat et al., 2021) (with notable exceptions e.g. Dennler et al. (2023); Ross et al. (2025)).

However, I agree with Talat et al. (2021) that failure to recognise one's positionality makes one complicit in harm, because (as I argue in this thesis) the harms of language technologies are the direct result of decisions made by those developing, deploying *and evaluating* these tools. Different disciplines, and indeed different authors within disciplines, use reflexivity and positionality to mean different things; I use reflexivity to mean the act of self-reflection on my own identity and biases, and positionality to refer to the declaration of my perspective as it compares to others, which is informed by my experiences of power and of oppression.

Positionality statement: A key aspect of my identity which informs my work looking at bias against LGBTQ+ identities is that I am myself a queer researcher. This gives me unique insights into harms and community tensions (e.g. I have lived experience of transphobia and biphobia from within and outwith the queer community). However, I am acutely aware of the limits of my own experiences; for example, I am not a queer person of colour nor a transfeminine person, two communities who are particularly vulnerable to violent harm. Whilst conducting my work (rather than purely as a retroactive reflection (Gani and Khan, 2024)), I attempt to address this lack of familiarity with academic literature and first-hand accounts. However my work is inevitably limited in its authenticity to communities outwith my lived experience, and my conclusions should be interpreted as such.

1.2 Scope

Social Bias Against Marginalised Groups Whilst bias can refer to any skew in a data set or the output of a model (e.g. sampling bias, statistical bias) in the following I use "bias" to refer specifically to *social bias* in model outputs. That is, the tendency of a model to encode and thus produce output that mirrors social stereotypes. This is typically measured by the relative probability of different outputs. This could mean differences in output token probability distributions given the identity term used in a prompt. It could also mean differences in accuracy of output across speaker groups (as I use it in Sigurgeirsson and Ungless (2024)), because the model has a poorer encoding of data related to a marginalised group and this is reflected in undesirable differences in the output. In the latter sense, I am encompassing some of what is referred to elsewhere as "fairness", though others have tried to disentangle these concepts (Cabello et al., 2023).

When I talk of bias in my own work, I always mean bias that harms a marginalised

community by encoding prejudice against this group (rather than for example "bias" against white men). In my work (and activism) I am focused on minimising harm to marginalised groups, rather than e.g. trying to ensure fairness through parity of treatment of marginalised and non-marginalised people (a distinction made obvious by analogy to affirmative action). This is driven by my understanding of societal power structures as systemic, meaning even if we build a "fair" technology, marginalised people will continue to suffer through exposure to an unfair wider system (the use context) (Miceli et al., 2022). Work that seeks to address demographic disparity in models but assumes a fair, meritocratic society has been labelled a form of "algorithmic idealism" (Davis et al., 2021); such works are "consistently eluded by the fairness they mean to achieve". Thus I maintain my focus on improving the lived experiences of marginalised people, which requires changing how the technology functions within that wider system, to the benefit of these people, an approach which Davis et al. (2021) refer to as algorithmic reparation.

Throughout this thesis, I use the phrase "social bias research" broadly, to encompass evaluating language technologies, be that low-tech solutions or state-of-the-art models, for social bias, fairness, or harms related to identity, within academia, industry or third-sector organisations; thus the intended audience for, and potential users of, my proposed approach is likewise broad. In places, I refer to those currently doing (or who might do) the work of social bias research as "practitioners". My work is also pertinent to third-parties using these technologies, who need guidance interpreting social bias evaluations.

Varied Publicly Available Tools My work is not specific to one particular model architecture, one particular use category (e.g. classification vs. generation), nor one particular task. I work across tools because (a) they are all impacted by similar issues (e.g. lack of (diverse) training data, toxicity in training data, oversight by developers when mitigating bias, complex power dynamics within the queer community etc.) and (b) this demonstrates the effectiveness of my proposed human-centric approach to bias research regardless of the specific technology.

My work primarily focuses on the harms of publicly available or public facing tools – such as commercial sentiment analysis systems, proprietary image generation systems and social media recommender algorithms – that may not be cutting edge, but likely have the greatest impact through the scale of their use. Indeed, some of the technologies I evaluate are not even ML-based – for example in Chapter 4 I consider word-list based sentiment analysis, and it is likely that the moderation system em-

ployed by TikTok, discussed in Chapters 6 and 7, relies in part on word-lists (Brown, 2021).

Specific Use Contexts As I ultimately argue in my thesis, measuring bias in abstract has limited benefits, my focus being the measurement of harms in context. To identify harms – not just bias – one must consider specific use cases (at least implicitly). Just as it is the dose that makes the poison, so too does the application of a technology determine its level of harm. To give an example from my own work (not included in this thesis), a speech synthesis system that can "mimic" queer voices is beneficial when used as an accessibility tool, and harmful when used on social media to mock queer people (Sigurgeirsson and Ungless, 2024): a system that can synthesise both queer and non-queer voices may be "fair", but it also has the potential to be harmful. This is an example of the issue of the "dual use" of NLP technologies highlighted by Hovy and Spruit (2016).

Queer Identities Work on bias in NLP has historically overlooked queer identities, except a handful of salient terms such as "gay" or "transgender" (with notable exceptions! e.g. Cao and Daumé III (2020); Dev et al. (2021); Lauscher et al. (2022) *i.a.*); I seek to address this oversight. Further, focusing on queer identities requires me to treat identity as fluid, non-binary, and open, expanding beyond much contemporary work on bias which relies on a simplistic, binary understanding of identity.

My work is inspired by intersectionality (Crenshaw, 1989) in that I consider how groups within the queer community are affected by the intersection of their queerness with other forms of marginalisation such as ethnicity, resulting in harms unique to this group.

English language in Western Context I limit my work to assessing bias in English language NLP technologies, which typically implicitly encode a Western context (as Bianchi et al. (2023) and Yu et al. (2022) found for text-to-image models, and Santy et al. (2023) found for NLP writ-large). This reflects my lack of familiarity with other linguistic and cultural contexts and is a limitation of all my work contained herein.

1.3 Structure

The content of my thesis is as follows:

- Chapter 2: I identify the myriad sources of bias in NLP technologies. I then sketch the typical approach to bias measurement and name five common lim-

itations Finally I summarise ethics, social science and HCI research that was influential to my approach.

- Chapter 3: I define my human-centric approach to social bias research in NLP (my five maxims for a more valid approach).
- Chapters 4 to 7: These can be thought of as case studies in this approach. Whilst the outline of my approach predates the first two studies (Chapters 4 and 5) it was through conducting these works that my approach was formalised.
  - Chapter 4: Queerphobic bias in commercial sentiment analysis tools
  - Chapter 5: Community response to transphobia in text-to-image (TTI) models
  - Chapter 6: "Folk theories" of perceived censorship on social media
  - Chapter 7: Creative solutions to perceived censorship on social media
- Chapter 8: In the conclusion I synthesise my findings and draw on scholarship by other researchers to argue that not only is a human-centric approach to studying bias beneficial, but that the act of measuring bias in abstract is pointless.

1.4 Key Contribution and Takeaways

My primary contribution is the development of a human-centric approach to social bias research with five key maxims. I intend this human-centric approach to act as a "counterweight" to the preponderance of techno-centric bias measurement work in the field (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023). Where the current typical approach is as follows, **my human-centric approach**:

- Focused solely on technology → **Sees NLP tools as part of large socio-technical systems**
- Focused on data bias → **Considers many sources of bias**
- Focused on an abstract, mathematical problem → **Focuses on the impact of technology on people in context - and how they respond**
- Based on authors' intuitions → **Is driven by social science theory and community knowledge**

- Focused on binary gender and race → **Addresses a broad range of demographics**

This change in approach necessitates a change in measurement style – from quantitative evaluations of model outputs, to mixed methods approaches to understand lived experiences of harm. It also necessitates a change in bias mitigation approach, from simple heuristics (which I show in Chapter 4 are ineffective and in Chapter 5 are unwanted) to forms of mitigation that take into account real human behaviour (which may include their own attempts at mitigation, see Chapter 7). Whilst proposing novel bias mitigation techniques is beyond the scope of this thesis, I hope my approach can be an inspiration to those working on this topic (as this approach has inspired my own work on bias mitigation elsewhere (Ungless et al., 2022)).

My thesis also contributes significantly to our understanding of queerphobia in NLP technologies. I demonstrate that both "low" and "high-tech" sentiment analysis tools can be biased against queer identity terms. I show that TTI models reflect reductive stereotypes about the trans and nonbinary community. Finally I identify that queer users of TikTok report disproportionate censorship.

In addition to proposing five maxims for more valid social bias research, and advancing our knowledge of queerphobia in NLP, key takeaways from my thesis are:

- Chapter 4: Queerphobic bias in commercial sentiment analysis tools
  - "Low tech" NLP models provided needed control
  - Superficial debiasing leaves less salient groups impacted by harm
- Chapter 5: Community response to transphobia in TTI models
  - Heuristic approaches to bias mitigation are widely rejected
  - "Obvious" solutions can still lead to harm
- Chapters 6 and 7: Response to perceived censorship on social media
  - Felt impact of an algorithm is the result of human-AI interaction
  - The public will develop their own "solutions" to perceived bias
- Overall:
  - Trying to prevent bias can directly cause harm
  - The public form complex beliefs about algorithms

Research on bias in NLP must shift from focusing on technologies in abstract to considering socio-technical systems, which requires studying human behaviour; without this, we will only ever develop a superficial understanding of the problem, which will lead to superficial solutions that fail to reduce harm.

In summary, my thesis contributes five maxims for human-centric bias measurement, along with compelling evidence that heuristic attempts to mitigate bias can introduce novel harms, and insight into the complex beliefs the public form around NLP technologies. It also contributes to the study of queerphobic bias in NLP, offering nuanced analyses of the harms of sentiment analysis tools, TTI models and recommender and moderation algorithms, to the queer community, grounded in social science scholarship.

Chapter 2

Background

In this Chapter, I begin by describing the many sources of bias in NLP technologies, vital context for understanding the limitations of existing approaches to bias measurement and mitigation. I then present critiques of the current typical approach, along five dimensions. Namely, I explain how bias evaluation work is frequently limited by being:

- focused solely on technology,
- focused on data as the source of bias,
- focused on an abstract, mathematical understanding of the problem,
- based on authors' intuitions,
- focused on a binary conceptualisation of gender and race.

Finally, I provide an overview of the ethics, social science and HCI research that has informed my own work. These three strands – the diverse sources of bias, the limitations of much contemporary evaluation work, and the rich scholarship on fairness, language and identity in other fields – fed into the development of my human-centric approach, which I set out in Chapter 3.

2.1 The Sources of Bias in NLP

Turn over any given rock in the woods and you will find insects underneath. Likewise, any NLP technology you care to test will reveal evidence of bias, be it for sentiment analysis, language of image generation, or for other language tasks (Zhao et al. (2017);

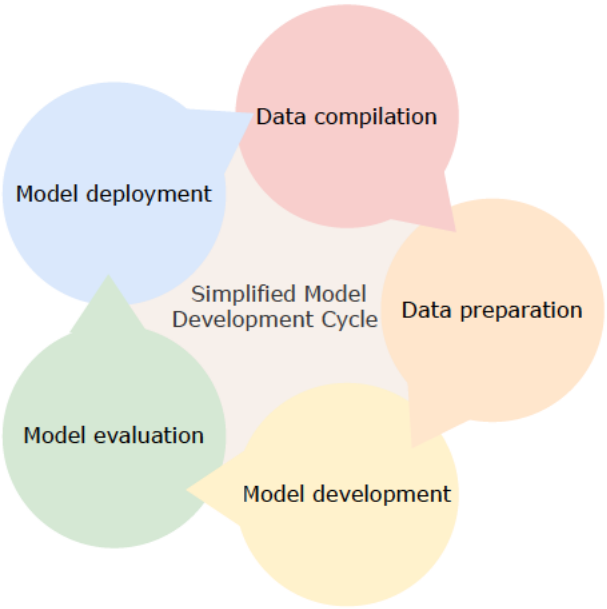


Figure 2.1: Simplified NLP model life cycle

Bolukbasi et al. (2016); Sheng et al. (2019); Zhao et al. (2018); Bianchi et al. (2023) *i.a.*). In the following, I will set out the sources of bias in NLP technologies that can ultimately result in them being harmful to marginalised communities. Crucially, I will highlight the role that human decisions play in determining model bias, which motivates my proposal for a human-centric approach to studying bias in NLP. As discussed in Section 1.2, I do not focus on a single architecture nor task, but draw insights from across NLP research (and beyond).

There are existing frameworks for categorising sources of bias in a technology: for example, Friedman and Nissenbaum (1996) propose pre-existing bias, technical bias and emergent bias; Suresh and Guttag (2021) propose a framework that distinguishes seven sources of harm in ML technologies, namely historical bias, representation bias, measurement bias, aggregation bias, learning bias, evaluation bias and deployment bias. In their work, Suresh and Guttag (2021) summarise the six stages of a typical ML system life cycle. Inspired by this work, I consider sources of bias at each stage in a simplified language technology life cycle, from data compilation and preparation through model development, evaluation and deployment, as depicted in Figure 2.1. This life cycle is greatly simplified: there will be overlap between these stages in terms of relevant literature, and in reality the stages will not progress linearly (for example, a model may be deployed then evaluated by third parties). The benefit of

this framework is that it allows me to emphasise the human decisions made at each stage of its life cycle which hugely shape model bias (Talat et al. (2021) refer to this as the embodiment of designers into the technologies). By reflecting on how choices made during development can introduce and reinforce biases in the system, I push the focus beyond imbalances in the data, which receives a significant amount of attention, but accounts for only part of the story (Hooker, 2021). As Hooker writes, "the overall harm in a system is a product of the interactions between the data and our model design choices" (Hooker, 2021).

One source of bias, which impacts all stages, is who is "in the room" (Táíwò, 2021) when models are developed - or more accurately, who is *not*. As noted in West et al. (2019), the overwhelming majority of AI developers and researchers are men and non-Black, and will be "blinkered" by their lived experiences.[1] The potential harm of some development decisions (e.g. the decision to use data from websites typically populated by men such as Reddit (Social, 2024) and Wikipedia (Schmahl et al., 2020); the decision to classify training data by perceived binary gender (Larson, 2017; Devinney et al., 2022; Sanchez et al., 2024) etc.) may thus be "invisible" to these developers. This blinkered life experience goes some of the way to explaining the potentially harmful decisions made throughout the project life cycle, even without the developers realising.

2.1.1 Data Compilation

The first stage of a typical life cycle is data compilation, a phrase proposed by Benjamin (2021) to emphasize the role that a developer[2] plays in actively selecting data. The selection process can introduce significant biases into a model.

A common focus of research into bias in NLP technologies is the role of data imbalances. That is, the imbalance between classes and labels, such that a label may be more strongly associated with a particular class due to its prevalence. Or the imbalance in co-occurence of certain identities and certain themes, for example sexual content featuring women. To give concrete examples, a lack of data about female doctors compared to male doctors may explain poor performance in coreference resolution (Zhao et al., 2018); an abundance of pornographic content involving women (Birhane and Prabhu, 2021; Birhane et al., 2021; Nichol, 2022) (especially transgender women, see Chapter

[1] A phrase used by interviewee D in Chapter 5 (unpublished data) to describe cisgender developers.

[2] Throughout this section I will refer to "a developer", but of course the vast majority of NLP technologies are the product of multiple teams of developers.

5) may explain sexually explicit output in TTI models for prompts related to women but not men (Wolfe et al., 2023). Data imbalances have been shown to be exaggerated in vision-language models' outputs (Zhao et al., 2017; Bianchi et al., 2023), although recent work by Seshadri et al. (2024) suggests this is primarily related to the use of unmarked language in prompts (though this is not terminology they use – see Section 2.2.2.4).[3] These data imbalances may reflect real world "imbalances" due to *historical bias* (e.g. which careers men and women were allowed to or chose to pursue)[4] or they may be a form of *representation bias* (the target population is not well represented in the data that the developer has chosen to use), per Suresh and Guttag (2021)'s distinction. Representation bias is particularly relevant to work on queer populations because they make up a small proportion of the general population (Ipsos, 2023), and due to social stigma and risk of harm may not explicitly "record" their queerness in data available to train models (Guyan, 2021, 2022; Sigurgeirsson and Ungless, 2024). Further, on the margins of society language around identities changes quickly,[5] and data from even a few years ago may no longer represent the queer community accurately (Guyan, 2022; Zimman and Hayworth, 2020). Imbalances in data lead to biased models that have learned "spurious correlations" (Schwartz and Stanovsky, 2022),[6] for example between gender & sexuality minorities and toxic content (Paullada et al., 2021). Hence a developer's decision to (re)use[7] a particular training data set is a source of bias.

Of course, in addition to the issue of demographic groups being represented in the data in an imbalanced way, there is the issue of certain demographics being entirely missing, so called "data gaps" (Markl, 2022; Criado-Perez, 2019). For example, data sets of voice data may be entirely missing data from gender non-conforming, non-cisgender or otherwise queer people (as I discuss elsewhere (Sigurgeirsson and Ungless, 2024)) (or those that are present may be "mislabled" by annotators, see Section

[3] An illustrative example: whilst there may only be slightly more images of male doctors than female doctors, there will be very few images of female doctors that do not include "female" in the caption, hence prompting with "doctor" alone tends to produce images of men (the unmarked gender). The topic of marked language is discussed in Chapters 4 and 5.

[4] It is up to the developer whether they wish to "correct" for such imbalances or simply aim to reflect current population data (Strengers et al. (2020)'s "adhering" vs "steering"), but this decision is often implicit, as I note in other work (Goldfarb-Tarrant et al., 2023).

[5] Innovation is also a particular feature of anti-languages, linguistic varieties generated by marginalised groups (Lefkowitz and Hedgcock, 2017; Halliday, 1976).

[6] Also referred to as "spurious associations" (Utama et al., 2020) or "questionable correlations" (Blodgett et al., 2020).

[7] The reuse of data (and of models) can create an algorithmic monoculture whereby individuals are harmed by algorithmic bias in a systemic way and inescapable way (Bommasani et al., 2022). See Thylstrup et al. (2022) for a further discussion of the ethical implications of data reuse.

2.1.2), just as they are missing data from certain global Englishes (Markl, 2022). Some marginalised communities, particularly marginalised queer populations, may have negligible presence in typical online data sources, or be hard to "reach" when compiling novel data sets, due to historical social exclusion or the risks of being publicly identified. As such, the final model is biased in that it is unable to accurately handle data from these "missing" populations, because of the (implicit) decision by the developer not to represent these communities.

Note that in the act of compiling data, there is (typically) a power imbalance between data source and developer – power as determined by knowledge (of the technology and its intended uses and risks), (financial) resources, and demographic group. This power imbalance raises ethical issues when compiling data from marginalised populations (Fussell, 2019; Mahelona et al., 2023), as we discuss in Chapter 5 with regards to image data from non-cisgender populations. Thus sensitively addressing data imbalances and missing data requires an approach that is cognisant of this power imbalance. Otherwise, the final model will be biased in that it is trained on data that does not authentically represent a community, but rather represents them through the eyes of the oppressor.[8] Examples of best practice from Indigenous Data Sovereignty scholarship can be found in Walter et al. (2021); Schwartz (2022); Bird (2020), where priority is given to authentically representing the world view of the community in the data collected about them (Walter et al., 2021). How the developer chooses to interact with the marginalised community – if at all – will impact the quality of the data they collect, and ultimately influences model bias.

Finally, the decision to use data which may be "algorithmically influenced" (a phrase I am using to refer to data that is selected by, generated by, or determined by the outcome of an(other) algorithm) can be a source of bias in the model output, particularly where this creates compounding feedback loops. For example, O'Brien (2012) writes of a "cascade of algorithmic bias errors" impacting Black defendants, who at each stage are more likely to receive harsher treatment, which then feeds into decisions made at re-arrest, impacting the training data and more deeply entrenching racial bias in the models. With regards to NLP technologies, one can imagine a scenario where a language identification (LID) model is trained on a manually selected subset of the data which an earlier biased model – which excludes texts in African American English (AAE) – had classified as likely being in English, which will entrench the bias

[8] A variation on the self-fulfilling prophecy discussed in Dwork et al. (2012) where data about the minority population are *deliberately* chosen to encode stereotypes into the model.

against AAE in the new model. Or otherwise, a TTI model may be further trained on data scraped from the internet which includes pairs of prompts and images which it had previously generated, which further entrenches the model's own biases (just as training on generated content degrades output quality (Hataya et al., 2023; Shumailov et al., 2024, 2023)). Beer (2022) talks of the data coil, a phenomenon whereby models that exist in a feedback loop are used in a way that generates new data which is used to train new model(s), which create similar feedback loops and ultimately entangle to form coils. Thus the decision to use algorithmically influenced training data (or at least tolerate its presence) can be a source of bias (amplification) in the model, which in turn influences future iterations and other models.

2.1.2 Data Preparation

After data has been compiled, it must be prepared for use in training a model, typically by "cleaning" the data (for example, to remove non-English text or to remove toxic content) and (depending on the training paradigm) by annotating the data (Gebru et al., 2020). Both of these preparation steps (cleaning and labelling) can introduce bias into a model.

A seminal work looking at how a typical step in data cleaning, LID, can introduce bias into training data is Blodgett and O'Connor (2017). The authors found that AAE[9] was disproportionately misclassified as non-English by several popular LID models. The resulting "clean" data would thus be a poor representation of AAE – and downstream models trained on this data would be likewise. Whilst there are LID models flexible enough to handle non-standard Englishes (Jurgens et al., 2017; Tan et al., 2020), it is likely that many recent downstream models trained on "clean" data continue to be biased due to lack of model "familiarity" with non-standard Englishes (the Englishes spoken by marginalised populations), largely because of the challenges of measuring dialectal variation and defining how much is permissible (Burchell, 2024).[10] The decision to clean data using a tool that is itself biased results in bias in the final model.

Another typical step in data cleaning[11] involves filtering out harmful content, be

[9] More precisely, data which was highly likely to contain AAE, which they refer to as African American aligned English data.

[10] See also Keleg et al. (2023)'s notion of "level of dialectness", which treats "dialectness" as a spectrum not a binary

[11] The phrase "clean" data evokes puritanical notions of moral cleanliness – the desire to "cleanse" can be a response to imagined contact with moral transgressors (Golec de Zavala et al., 2014), see also

that sexually explicit content or hate speech. For marginalised populations, this presents a double-edged sword: for example, filtering out sexually explicit content (to reduce the likelihood of harmful, typically misogynistic output) might introduce gender imbalances in the data that actually results in biased models, as Nichol (2022) reports for the TTI model DALL·E 2. With regards to hate speech, if this were not filtered, the amount of toxic content related to marginalised communities would be significant (Luccioni and Viviano, 2021; Maronikolakis et al., 2022), leading to the issues associated with data imbalances discussed above. However, filtering out data related to queer terms can also be harmful. Bender et al. (2021) note for example that the Common Crawl corpus, a widely used text data set for training NLP technologies, is cleaned of content containing queer identity terms such as "twink", which they argue will "attenuate... the influence of online spaces built by and for LGBTQ people". The list of words[12] used to clean this data also contains reclaimed slurs such as "fag", "bitch" and the n-word,[13] and words likely to be associated with queer content such as "sexuality". Thus a model trained on "clean" data may be biased in its (in)ability to handle input relating to marginalised (e.g. queer) identities, as a result of the (implicit) decisions by the developer on what constitutes appropriate training data.

Whether to retain reclaimed slurs requires careful consideration, taking into account the intended use of the data and resulting models; however, the approach to data *collection*[14] and preparation that sees data as a resource to be mined/ scraped and seeks to collect as much as possible, to be used in as many applications as possible, leaves little room for such nuance (Thylstrup and Talat, 2020; Benjamin, 2021; Paullada et al., 2021). Decisions on what counts as "desirable" content are likely not fully interrogated by "blinkered" developers. This leave room for a developer's implicit biases to impact the resulting training data. For example, the decision to remove sexually explicit content leaves room for the developer's own sexual preferences to bias the filtering models they develop, even without their awareness (Gehl et al., 2017) – just as the list of obscene words discussed above reflects what is culturally salient to its authors, being far from comprehensive. A developer will also often have commercial motivations to filter out "undesirable" content - we return to the topic of commercial motivations in Section 2.1.3. A developer's implicit or explicit preferences will result in model bias

Zhong et al. (2010).

[12] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

[13] Most toxicity detection systems, be they word lists or ML systems, return false positives for reclaimed slurs (Röttger et al., 2021).

[14] See Benjamin (2021) for a discussion of *collection* vs. *compilation*

where they align with power structures, such as the marginalisation of queer identities, as will often be the case given the typical demographics of AI developers (discussed above) (West et al., 2019).

Beyond "cleaning" of data, a typical data processing step involves labelling data. This can be a significant source of bias, as labels can reflect the developer's and annotators' (implicit) social biases, and even be overtly offensive (Crawford and Paglen, 2021). As Paullada et al. (2021) highlight in their survey of data set development in ML, the act of annotating data is a form of "interprative" work, and this is evident in both the ascription of alt-text to images (which we discuss below), where the "annotator's" goal is unlikely to be model development, and in the ascription of labels to data for supervised training. Annotators' own biases will inevitably influence their choice of label (Excell and Al Moubayed, 2021; Al Kuwatly et al., 2020), just as the developer's biases and intuitions will influence their definition of a labelling scheme (Crawford and Paglen, 2021; Bennett et al., 2025). Even where detailed annotation guidelines are provided to try and prevent annotator bias from impacting labelling decisions, these attempts can be hampered by cross-cultural differences (Smart et al., 2024) (and are limited by the developer's own imagination).[15] Further, such annotation guidelines are rarely made public, making it impossible for third parties to critique the labelling scheme, clarity of instructions etc. (Paullada et al., 2021). Where majority voting is used to resolve differences, minority marginalised voices may be "silenced" (Abercrombie et al., 2024), even though their lived experience likely makes them a better judge of e.g. whether content is authentic, offensive etc. (though cf. Abercrombie et al. (2024) who did not find gender to be a predictor of misogyny labelling). Thus annotator biases may result in biased models (cf. Miceli et al. (2022) for the limitations of focusing on "worker bias"), despite attempts to prevent this, and the developer's decisions when defining the annotation scheme and resolving disputes can also introduce bias.

Issues may also arise from the use of inappropriate or oversimplified proxies (a form of measurement bias, per Suresh and Guttag (2021)'s framework). For example, the decision to label data with a single, static gender label fails to take into account the fluid nature of gender (Guyan, 2022), and may inappropriately conflate gender identity, gender presentation and perceived gender (Devinney et al., 2022; Sanchez et al., 2024).

The act of pairing an image with its alt-text for model training can also be thought

[15] In Chapter 5, we specifically complement our analysis of model outputs with a survey in order to go beyond the harms we could envisage *prima facie* to include in our annotation scheme

of as an act of labelling. Alt-text serves many purposes, including search engine optimisation (SEO) and accessibility (Hong et al., 2024). Alt-text will capture the image subjects' perceived gender, sexuality, ethnicity etc. which may not reflect their identity. For example, it is likely that images of queer people at Pride are taken and "labeled" by strangers, often people from outwith their community, who may through ignorance or malicious intentions label the data in an inaccurate or even offensive way.[16] Thus the words associated with (imagery of) the identity are often the words of the oppressor, and the model's association with the identity are the oppressor's associations. For example, images of indigenous Americans may be shared by outsiders precisely because they play into stereotypes of how indigenous people live, hence I found TTI models were incapable of producing images of Two-spirit people at work conferences (see Chapter 5), because traditional corporate jobs are not associated (by non-indigenous people) with indigenous people. Thus the biases of the "annotators", and the developer's decision to rely on these inherently subjective image descriptions, result in model bias.

2.1.3 Model Development

There are a number of ways that model design and training can introduce bias into a system, both through the choice of *what* to model, and the choice of *how* to model it. Shah et al. (2020) write that "bias is not necessarily something gone awry, but rather something nearly inevitable in statistical models." By virtue of being designed primarily to detect and replicate patterns in training data, models "inevitably" mimic patterns of discrimination.

Considering first model design (the "how"), specifically loss function design, as Hooker (2021) writes, whilst "standard loss functions do not explicitly encode preferences for other objectives we care about such as algorithmic bias, robustness, compactness, or privacy... [this] does not mean they have ceased to exist". A loss function designed to optimise performance on the bulk of the data will result in relatively poor performance on "the long-tail", which often relates to marginalised, minority[17] identities (Hooker, 2021). This discrepancy is called "learning bias" in Suresh and Guttag (2021)'s framework. Thus loss function design can introduce bias into a system by

[16] Using Elazar et al. (2024)'s application programming interface (API) I found thousands of examples across six popular corpora of the *n-grams* "transgendered woman" or "trans identified male", a non-recommended and offensive way to refer to trans women, respectively.

[17] In the population size sense.

implicitly de-prioritising the data of marginalised groups.

As my thesis encompasses many different models and my findings are intended to be somewhat task agnostic, I will not explore in detail how specific model architectures are linked to increased bias. However, I will briefly note that for language models, increasing model size does not seem to directly address bias (Tal et al., 2022) (though cf. Utama et al. (2020)), however, there is evidence that compression through e.g. pruning and distillation increases bias in vision models and language models (Silva et al., 2021; Hooker et al., 2020). Thus it is the case that, at least for language and vision models, choice of model size can be a source of bias.

Whilst many developers attempt to address bias through model design and finetuning,[18] this is frequently ineffective (Gonen and Goldberg, 2019)[19] and can even introduce new biases (reminiscent of the impact of data filtering). Xu et al. (2021) found that debiasing language models to decrease the likelihood of toxic content also decreased the likelihood of the model producing content which mentions minority identity terms. Thus the decision to try and debias a model (e.g. reduce disproportionate hateful responses related to marginalised identity) can actually introduce novel biases (e.g. erasure through the low likelihood of certain identity terms).

With regards to the choice of what to model, Miceli et al. (2022) write that "machine learning systems are fundamentally trained to cluster and classify data. When these classifications are value-laden and interest-informed, they result in imposing and promoting the... worldviews of some groups" – hence the choice of what to classify, what to model, is a source of bias. Further, just as Cobbe (2021) writes of algorithmic moderation tools allowing the insertion of "commercial considerations into everyday communications", so too do proprietary NLP technologies allow their creators' commercial priorities to influence a vast array of outcomes. For example, generative technologies like ChatGPT and DALL·E 2 insert Open-AI's "commercial considerations" (a desire for mass-palatability) into a broad range of tasks including the creation of art etc. (a diversity of use cases which they encourage!). Commercial considerations may favour "neutrality" over showing support for a marginalised community, arguably an example of bias against the marginalised community, as was found by journalist Mona

[18] There is limited evidence for the efficacy of debiasing "upstream" i.e. debiasing a language model that will ultimately be finetuned for e.g. sentiment analysis, hate speech detection, question-answering (QA) etc. (Steed et al., 2022) (cf. (Liang et al., 2020)) - see the discussion in Section 2.2.2

[19] Particularly due to a lack of relationship between intrinsic measures, typically used to optimise models in development, and extrinsic harms (Cao et al., 2022; Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022)

Chalabi for ChatGPT in response to questions about Palestine compared to Israel.[20] Further, a model may be released that remains biased against a known (or unknown) list of identities because these communities are not commercial priorities.[21] Commercial considerations can be the source of bias (going unaddressed) in the model.

Finally, training a model may legitimise the decision to try and predict certain traits. Paullada et al. (2021) write that the collection of data and training of models to predict such "fluid, subjective personal traits... presuppose[s] that these predictions are... worthwhile to make". This can afford an air of objectivity and neutrality to a highly political decision (Talat et al., 2021). Thus the very decision to create a model to predict e.g. sexuality from voice data, or gender from name data, can be harmful, as critiqued by Sigurgeirsson and Ungless (2024) and (Gautam et al., 2024) respectively. Further, bias is likely to occur as these models inevitably rely on "meaningless shortcuts" to predict qualities such as sexuality (Paullada et al., 2021). Because these models typically rely on spurious correlations or stereotypes e.g. that a particular author is male because in a Western context their name is more likely to be given to men (Gautam et al., 2024), their output is likely to be biased.

2.1.4 Model Evaluation

The effective measurement of bias in NLP technologies being the primary focus of this thesis, I return to this topic in depth in Section 2.2. Here I will briefly note some ways in which the evaluation of NLP technologies can introduce bias. Whilst I present model evaluation before deployment, in reality much evaluation happens after public release through community red-teaming, auditing by third parties, and the public's everyday algorithmic audits (see Chapter 7) – often relying on work by marginalised communities to protect themselves (Buolamwini and Gebru, 2018).

Prioritisation of one particular performance metric at the expense of metrics related to bias could be thought of as a source of bias in the model. Just as Hooker (2021) writes of the design of loss functions de-prioritising other objectives, so too does the decision to evaluate on overall F1-score, precision, BLEU etc. de-prioritise minimising bias. Thus the model is biased because it has been trained to optimise overall performance and it (likely) relies on "spurious correlations" (Schwartz and Stanovsky, 2022) (e.g. between queer identity terms and hate speech labels) to achieve this.

[20] https://www.instagram.com/monachalabi/p/CydbE5sutDQ/

[21] Akin to the infamous story of an Apple employee being told Siri did not work for AAE because Apple products are for a "premium market" (Benjamin, 2019)

Evaluation benchmarks can similarly be a source of bias in models, where they prioritise performance on data from, by or about non-marginalised groups. For example, if data from a marginalised group is not included in the benchmark test set or makes up only a small proportion of the data, this means that scores will be relatively unaffected by poor performance on this group. Effectively this means models are not sufficiently "punished" for being biased. The well-known Gender Shades paper demonstrated bias against women of colour in popular facial recognition benchmarks (Buolamwini and Gebru, 2018). Unless benchmark results are differentiated across marginalised identities, this bias remains invisible – one could say hidden. Hence (optimising for) evaluation benchmarks can be a source of bias in a model.

Further, even (optimising for) *bias* benchmarks can be a source of bias, where the benchmarks themselves further encode social biases. Bias benchmarks typically focus on a small number of salient demographic features (as I have shown elsewhere for language generation (Goldfarb-Tarrant et al., 2023)), failing to account for less salient, often greatly marginalised groups. I return to this topic in Section 2.2.2.5. For example, to my knowledge there are no benchmarks which measure bias against sex workers or people living with HIV/AIDS, all of whom face significant discrimination and all of whom are likely to be affected by model bias. These test sets can then be thought of as a source of bias in the model where they influence the developer's priorities. For example, a benchmark that tests for bias against gay and transgender people may encourage model developers to create a model which rarely mentions sex work or HIV/AIDS, to avoid stereotyping[22] such as was demonstrated in Nozza et al. (2022). However, that model would then be guilty of erasure against sex workers or those with HIV/AIDS (this parallels the erasure of marginalised identities by detoxified models discussed in Xu et al. (2021)). Further, the choice of bias metric can be a source of bias: Blodgett et al. (2021) found that benchmarks for language generation models (as well as being poorly constructed) could "[encourage] models to produce stereotypes just as often as anti-stereotypes" through the use of aggregated metrics which assume stereotyped and anti-stereotyped phrases should be equally likely, which may be undesirable. This also fails to account for confirmation bias:[23] output which aligns with existing human biases will be favoured during information processing, so be more memorable – more potent. Optimising for performance on such a benchmark

[22] See Nadal et al. (2014); Anzani et al. (2024); Rice et al. (2022) for discussions of stereotypes associated with gay and transgender people.

[23] Elsewhere, I discuss the failure to acknowledge confirmation bias for language generation evaluation (Goldfarb-Tarrant et al., 2023).

results in a model that can be said to be biased, at least through the lens of confirmation bias.

In conclusion, evaluation metrics and benchmarks, or more accurately the decision to optimise for final performance on these tests, can be a source of bias where they fail to incorporate marginalised groups sufficiently, if at all, and where the choice of aggregate metric fails to account for power imbalances.

2.1.5 Model Deployment

Decisions made when deploying a model can be a direct source of bias in the model's output, as well as biasing outcomes based on the model's output. Selbst et al. (2019) highlight five traps that can impact the fairness of an ML system, and I discuss here two that can be sources of bias at deployment. The Portability Trap – the desire to reuse models for new tasks and in new contexts – can be a source of bias in a model's output where performance varies by identity in the new population. For example, a spoken LID model trained to identify English primarily using data from southern England may be biased against working class people when used to identify English from speakers elsewhere in the UK. This is because working class people are more likely to have "regional" accents and incorporate elements of local dialects, and so will diverge more from the training data than middle class people (Hughes et al., 2012).[24] Hence the decision to reuse the model for a new population becomes a source of bias in the model's output.

Another trap Selbst et al. (2019) identify is the Framing Trap - that is, failure to account for all entities in a socio-technical system. They give the example of a situation where a judge systematically ignores the recommendations of a recidivism evaluation model that has been "de-biased" such that biases (for example, against men of colour) are re-introduced.[25] To give an example from NLP technologies, a system for screening resumes may be said to be unbiased, but *outcomes* may nonetheless be biased if the recruiter tends to ignore the recommendations in a systematic way (even unknowingly). Whilst these are not examples of model bias per-say, they highlight that measuring model bias in abstract can be a poor proxy for harm. The models can add an air of objectivity (as discussed above) and legitimise decisions made my humans (such

[24] In general, social class is a much overlooked demographic factor in NLP research (Cercas Curry et al., 2024).

[25] Just as the public ignore recommendations from a fair career recommender model in favour of those that align with their existing stereotypes (Wang et al., 2022).

as judges, recruiters etc.). Just as a treatment plan cannot be said to be effective if it is known patients will not stick to it, a model is not really "de-biased" if it is likely users will ignore outputs in a systematic way (e.g. favour upper bound recommendations for one group and lower bound for another).

Finally, when a developer chooses to include some kind of filtering at inference (either ML- or n-gram-based), such as when the model "refuses" to respond to prompts containing queer identity terms, as I note in Chapter 5 (and as has been found for e.g. the Spotify playlist creation AI,[26] Bard (Fredriksson, 2024) and Copilot,)[27] or gives some variation on a canned response (as is evident with ChatGPT when prompted on trans rights), or conducts filtering on its own output (as we found for Stable Diffusion in Chapter 5), this can introduce bias into the model's output. This is because it is likely that marginalised identities – strongly associated with toxicity in the training data and filtering models due to historical bias – will be subject to additional moderation compared to non-marginalised identities. Hence the decision to rely on filtering at inference for "safety" can be a source of bias in the model's output. We return to this topic of post-hoc filtering in Chapter 5.

2.1.6 Summary

In this section, I have highlighted how human decisions at every stage in a model's life cycle can be responsible for bias in its output. **Ultimately, it is human behaviour that determines bias in NLP technologies**. Whilst this behaviour is inexorably entangled with (the output of) these technologies, and bias can be said to be the product of this entanglement, I focus on humans as determining bias because it is humans who choose which technologies to build, where and how to deploy them, how to interpret the output etc. When one considers the number of stakeholders making interlinked decisions that are involved in developing an NLP model, it is clear that addressing bias is almost intractably complex. It is not as "simple" as improving data compilation to be more balanced, because the cleaning of this data can (re)introduce biases; it is not as "simple" as improving the annotation guidelines to avoid annotator bias, because the loss function may introduce reliance on "spurious correlations" leading to bias; it is not as "simple" as choosing the model which performs best on bias benchmarks, because the benchmarks themselves may be biased; it is not as "simple" as deploying

[26] https://x.com/jkrwls/status/1787982145924841926

[27] When working on our paper on synthesis of "gay voice" (Sigurgeirsson and Ungless, 2024), my colleague found Copilot would not complete code which included the word "gay"

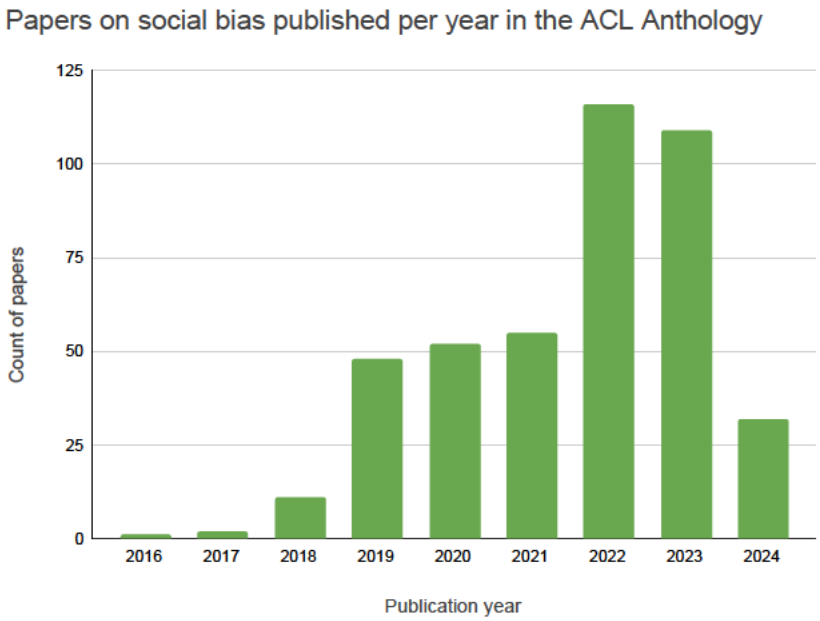


Figure 2.2: Chart showing the number of papers on social bias in the ACL Anthology each year, a rough proxy of the scale of interest in the topic within the NLP community.

a debiased model, because new use contexts will introduce new biases. Likewise, the measurement of bias at a given stage of development (e.g. measuring data imbalances, bias upstream before finetuning etc.) paints only (a sometimes conflicting (Goldfarb-Tarrant et al., 2021; Cao et al., 2022)) part of the picture. In the following, I present and critique the typical approach to measuring and mitigating bias in NLP, which often treats bias as a technical issue and fails to account for the human factors discussed above. I then summarise the literature from other fields which shaped my thinking. In Chapter 3 I present my own approach which, in response to the role of human decisions in determining model bias – and to a historical preoccupation with the technological aspects of bias (Blodgett et al., 2020) – is human-centric.

2.2 Measuring and Mitigating Bias in NLP

2.2.1 The Norm

Between 2016 and 2022 the number of papers on bias measurement and mitigation in NLP increased almost exponentially:[28] Figure 2.2 visualises the number of papers

[28] I postulate that this exponential growth was only tamed by a shift in language from "bias" to "safety", with related work still being plagued by the same issues of lack of clarity and lack of validity.

looking at social bias in the ACL Anthology 2016-2024.[29] Well-cited examples across different NLP tasks include Bolukbasi et al. (2016) who find stereotypes encoded in static word embeddings; Zhao et al. (2017) who find image classification models make systematic errors that actually exaggerate data imbalances with regards to gender; Zhao et al. (2018) who find co-reference resolution models make systematic errors that align with career/gender stereotypes; Sheng et al. (2019) who find language generation models produce stereotyped content about marginalised identities; Zhang et al. (2020) who find toxicity classifiers tend to return false positives for sentences containing demographic terms; Dhamala et al. (2021) who quantify bias in language generation models by measuring (amongst others) toxicity and gender "polarity" in response to prompts containing identity terms; Smith et al. (2022) who measure bias in language generation models against around 600 identities through a range of metrics including level of (inappropriate) sympathy; Bianchi et al. (2023) who find TTI models produced stereotyped images in responses to prompts about different human qualities and careers; and Dev et al. (2024) who, through community engagement, create a resource for testing language models for diverse stereotypes specific to India. Such papers were vital in drawing attention to the issue of model bias which could, when models were deployed in real-world scenarios, lead to harm against marginalised communities.

Typically, measuring bias in NLP technologies involves identifying an affected demographic group; creating a test set of prompts, (labelled) sentences or image-text pairs etc.; defining a quantitative metric of bias; and comparing scores across models, hyperparameter settings etc. The prevalence of this approach is, I feel, uncontroversial: the reader would be challenged in finding papers which do *not* follow it. Choice of demographics and the way these demographics are indexed in the test data are often based on authors' intuitions or terms borrowed from previous bias research papers (Goldfarb-Tarrant et al., 2023; Blodgett et al., 2020). Evaluation typically relies on a automated quantitative metric such as difference in accuracy e.g. Röttger et al. (2021), log likelihood e.g. Nangia et al. (2020), percentage of "toxic" content as determined by a toxicity classifier e.g. Chowdhery et al. (2023) etc. Any qualitative evaluation often relies on a small handful of outputs, as Calabrese et al. (2021) find for the evaluation of bias in hate speech detection.[30] Bias evaluation often happens without reference to specific use contexts, as is typical of ML evaluation (Hutchinson et al., 2022; Goldfarb-

[29] Papers were selected if abstract included ["bias"] and any of ["soci[oe]*","social bias", "harm*", "identity", "demographic", "gender","ethnic*","disab*","sex*","soc* class","appearance","nation*"]

[30] And as has been noted for performance evaluation in explainability methods in NLP (Nauta et al., 2023).

Tarrant et al., 2023).

This approach to bias evaluation has been valuable in that it is efficient and relatively low-cost, and the incredible speed of NLP development demands evaluation work keeps pace. Many of the papers cited above were pivotal in drawing attention to the issue of bias in novel technologies. By defining a pragmatic scope, practitioners can more quickly ring the alarm on "blatantly sexist" models (Bolukbasi et al., 2016) – and other forms of bias – and iterate over mitigation strategies. Quantitative evaluation resources like WinoBias (Zhao et al., 2018) and SPICE (Dev et al., 2024) offer an estimate of model harm that allows for the quick comparison of different architectures, hyperparameters etc. However, work that adopts this approach is often hampered by one or more of the five limitations which I set out below.

2.2.2 Critiques

2.2.2.1 Focused solely on technology

Work that adopts the current typical approach to bias measurement and mitigation often does not account for the social aspect of bias. Bias in NLP is a socio-technical issue (as Dolata et al. (2021) discusses with regards to algorithmic fairness), in the sense that it is the product of the interdependence between human behaviour and model output (Dolata et al., 2021) (the entanglement I refer to above). Models are biased precisely because "the algorithm creation process is a social practice" (Dolata et al., 2021), social practices which might be themselves influenced by data and algorithms in recursive coils (Beer, 2022). The influence of human- and of technological- factors on a model's output cannot be neatly separated. Bias metrics are likewise socio-technical because they are the result of the (re)translation between human concepts (such as "offensiveness"), mathematical formulations, and interpretation of mathematical results (a translation process that is often left implicit, as I have argued elsewhere (Goldfarb-Tarrant et al., 2023)). When studying bias in NLP, focusing solely on the technology may be pragmatic, as it allows for a more clearly defined scope. However, it also offers a more limited understanding of the problem, and crucially misses many opportunities for improvement. We must consider the socio-technical systems these technologies are embedded within.

2.2.2.2 Focused on data as the source of bias

A number of issues follow from the failure to account for the socio-technical nature of bias. A predominantly technical approach may overlook the role that human behaviour plays in determining bias, which means certain sources of bias go unnoticed. For example, Hooker (2021) argues that much work on bias ignores the role of model design decisions in determining model bias. She argues that a pre-occupation with data imbalance "invites diffusion of responsibility" which absolves developers of responsibility for how their design decisions determine model bias (see Section 2.1.3). Likewise, an approach to bias measurement and mitigation that focuses solely on data imbalance will fail to account for the influence of developer behaviour. Training data is undeniably a significant contributor to model bias, and addressing data imbalance is vital. However, focusing solely on data is a problem because overlooking sources of bias again means overlooking opportunities for improvement – like trying to understand what went wrong with a cake by only examining the ingredients and not the order they were combined, mix time, oven temperature etc.

2.2.2.3 Focused on abstract, mathematical problem

Operationalising bias as a metric that can be measured efficiently at scale allows practitioners to provide swift estimates of the potential harmful impact of a (new) model. For example (in addition to those given above, such as Zhao et al. (2018) and Dev et al. (2024)), Marchiori Manerba et al. (2024) propose a probing framework which uses perplexity to measure the relative association between hundreds of identities and thousands of stereotypes, a scale not feasible with qualitative approaches.

However, this technical approach to bias, which encourages the treatment of bias as something that can be quantified and thus *optimised for*, can also lead to a false sense of fairness (Davis et al., 2021), and a false hope in the effectiveness of debiaising techniques. This mirrors the notion of "algorithmic idealism" from Davis et al. (2021), who write that computational approaches which see the goal of fair ML to "neutralize" demographic disparity will "consistently fall short" because they fail to account for the entrenched nature of discrimination. For example, Gonen and Goldberg (2019) find that a popular debiasing technique that involves neutralising the gender projection of non-gendered word vectors (such as career terms and human qualities) – that is, their similarity with gendered terms such as "mother", "he" – is "superficial", and simply acts to hide the bias in other dimensions. Similarly, Bianchi et al. (2023) find that

attempts at debiaising DALL·E through data filtering and balancing techniques, and the inclusion of guardrails at inference time, are ineffective at reducing many forms of bias. Elsewhere I discuss the pitfalls of trying to measure bias in language generation models through an inappropriate, but easily quantifiable, proxy such as sentiment score (Goldfarb-Tarrant et al., 2023); training a model to minimise discrepancies in average sentiment score may do nothing to address harmful stereotypes in model outputs.[31]

Operationalising bias as a quantifiable metric is valuable for at-scale evaluation; however, by attempting to isolate a single metric or set of metrics, this can offer only a superficial understanding of an issue which is in fact deeply rooted in the entire socio-technical system (Davis et al., 2021). I argue that a superficial understanding of bias as a purely technical issue – a mathematical kink in the model – will always result in superficial solutions, that may offer some improvements, but which ultimately fail to fully address the issue. The same solutionism that leads developers to believe all things can be modelled (Selbst et al., 2019), leads developers to assume all issues can be quantified, even concepts such as "ethicality", which LaCroix and Luccioni (2022) argue defy quantification. Automation and quantification are relied upon to provide authority and legitimacy to AI ethics work (Widder, 2024), and this applies likewise to NLP bias research. Further, the Grey Hoodie project (Abdalla and Abdalla, 2021) explicitly calls out the role that this focus on technology as "a mathematical problem" can play in helping developers avoid some of the concerns raised by those outside Big Tech.

2.2.2.4 Based on authors' intuitions

Those developing and researching language technologies often rely on their own intuitions when measuring and mitigating bias. Once again, a motivating factor is the pressure to keep pace with the deployment of new models in new contexts. Interdisciplinary work is time consuming, and many institutions are not well set-up to facilitate it (Tobi and Kampen, 2018; Fischer et al., 2011). Thus, work often fails to incorporate social science and/or linguistic research (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023), nor community input (with notable exceptions e.g. Dennler et al. (2023); Dev et al. (2024); Gadiraju et al. (2023) *i.a.*). Failing to draw on relevant expertise and

[31] To give a concrete example, a model may produce equal numbers of positive sentences about transgender and cisgender people, but if the positive sentences about trans people always make reference to bravery and self-actualisation, and the sentences about cisgender people never make reference to identity, this model is still reflecting benevolent cis-sexism in its output. I discuss models producing "inspiration porn" about transgender people in Ungless et al. (2025a).

experience means that NLP bias research can be guilty of imprecise use of terminology leading to lack of comparability; inefficiently "re-inventing the wheel"; failing to account for how language shapes identity; and being blinkered by the developer's own (limited) experience.

With regards to a lack of precision, terms such as bias and fairness are used interchangeably, despite the fact that they can be orthogonal: a debiased model could in theory hurt performance for certain demographic groups, meaning the model is debiased in its representation but unfair in its outcomes (Cabello et al., 2023).[32] Blodgett et al. (2020)'s seminal paper on the topic of bias research in NLP highlights a lack of precision in the use of the term, and a disconnect between bias metrics and real world harms - critiques of the field we found to still hold three years later despite the ubiquitous nature of the 2020 paper (Goldfarb-Tarrant et al., 2023).

With regards to failing to account for social science research, Bianchi et al. (2023) find that attempts at debiaising DALL·E are ineffective because they fail to mitigate biases that are not salient in computer science bias scholarship, such as the stereotyping of African people as poor. The developers of DALL·E, in relying on a limited understanding of bias based on other computer science scholarship, thus fail to mitigate bias against the majority of identities. Blodgett et al. (2021) find that four popular language model and co-reference resolution bias benchmarks include comparisons that have no bearing on societal power dynamics e.g. comparing performance on sentences about American vs. Scottish horses. Models may be "biased" against Scottish horses, but it is not clear this leads to any harm.[33] Elsewhere I encourage researchers not to "re-invent the wheel" and consider relevant social science research when developing bias metrics (Goldfarb-Tarrant et al., 2023), including research on power structures.

With regards to a lack of linguistic research, Blodgett et al. (2020) argue that current approaches to bias measurement fail because they do not account for "the role that language plays in maintaining social hierarchies".[34] Researchers typically rely on identity terms, pronouns or names to measure bias (Goldfarb-Tarrant et al., 2023; Gautam et al., 2024), and do not consider the many ways identity is encoded in language e.g. dialect, word choice etc. (with exceptions e.g. Blodgett and O'Connor (2017); Zhang et al. (2021) *i.a.*). This means many types of bias go overlooked. I have found

[32]The model would still be considered biased in my use of the term.

[33]I would not consider these models biased as this does not reflect social stereotypes.

[34]Failure to incorporate linguistic expertise is an issue that extends well beyond just bias measurement in NLP, for example Markl and Lai (2023) highlights a lack of linguistic precision when using terms such as "accent" and "native speaker" in speech technology research.

numerous examples of bias mitigation work that do not incorporate relevant linguistic scholarship; I will give two examples to illustrate my point. Seshadri et al. (2024) demonstrate that many reports of data imbalance exaggeration in TTI models can actually be accounted for by the fact that typically gender is only mentioned when it goes against a stereotype. That neither they, nor the authors they critique, make reference to the theory of marked identity and marked language (Bucholtz and Hall, 2004) suggests a lack of integration of linguistic scholarship. Likewise in An et al. (2022), who use dictionaries as "objective, impartial, and concise definitions of words" and "unbiased reference points": this fails to account for (meta)lexicographical scholarship on the subjective nature of dictionaries (Łozowski, 2017). Incorporating linguistic expertise would ensure bias measurement and mitigation approaches account for how language is used to create our identities (Bucholtz and Hall, 2005). For example in Chapters 4 and 5 I compare marked identities (e.g. transgender) to implicit and explicit norms (i.e. where the typically unmarked identity is "spelled out" e.g. "cisgender").

Finally, bias evaluation often fails to take into account community knowledge (Dennler et al., 2023).[35] In their review of what is often overlooked in AI ethics research, Birhane et al. (2022b) find that papers rarely account for the lived experiences of those facing marginalisation. Incorporating community knowledge is vital if researchers and developers want to make a tangible difference to the lives of marginalised people.

2.2.2.5 Focused on binary gender and race

Bias work in NLP tends to focus on (binary) gender and racial bias in a way that relies on a simplistic understanding of both of these constructs, and to the exclusion of considering other marginalised groups (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023).[36] Typically, gender is presented as a binary construct (Devinney et al., 2022; Larson, 2017; Sanchez et al., 2024), and work fails to make a distinction between perceived gender, gender presentation, gender identity, sex and/or grammatical gender. This results in issues such as inclusion of "intersex" in a list of "gender neutral words", or using perceived gender and sex interchangeably (both of which I have seen in manuscripts on `Arxiv`).

[35] An issue that affects other areas of NLP research such as hate speech detection (Maronikolakis et al., 2022)

[36] This pre-occupation with gender and ethnicity also impacts hate speech data sets (Yu et al., 2024).

2.3 Insights from Other Fields

In order to shift the focus of bias evaluation work from the technological to the human, it is vital to integrate scholarship on fairness, language and identity from other fields, lest we "re-invent the wheel". Herein I briefly summarise work that was influential in the development of my approach. I treat bias as a systemic issue (Feagin, 2013; Feagin and Bennefield, 2014), something that is embedded into systems including technologies. This means structural change is required to reduce disparities (rather than e.g. the removal of bad faith actors). Davis et al. (2021) critique the algorithmic idealism that attempts to achieve fair ML through computational methods which ignore that discrimination is systemic. They call for bias to be addressed through the lens of equity, not equality, whereby people are allocated the resources needed to achieve equal outcomes in a process they refer to as "algorithmic reparation" (Davis et al., 2021); I adopt this focus on equity over equality in my own work.

The marginalised communities themselves will be best positioned to name what resources – technological or otherwise – are needed to enable equal outcomes. HCI offers several frameworks for designing new technologies which embed these perspectives, for example Value Sensitive Design (Friedman, 1996; Friedman et al., 2013) and Participatory Design (Sanders, 2002). I am inspired by these design frameworks to explicate the values embedded in the technologies I evaluate and the resources I create, and to involve affected communities in the evaluation process. My work is also inspired by the Indigenous Data Sovereignty movement, which declares that data about a community should be "inclusive of wider social structural context" and reflect a "nuanced narrative" of the community (Walter et al., 2021).

Fundamental to my understanding of language and identity are the notions of indexing (Bucholtz and Hall, 2005), and of marked identity (Bucholtz and Hall, 2004). Bucholtz and Hall (2005) explain that word choice, stylistic features and more contribute to the *production* of identity in conversation, which they refer to as indexicality. Markedness refers to the phenomenon whereby powerful identities, such as whiteness, are assumed as the norm (unmarked), whereas non-normative (marked) identities are associated with linguistic practices (such as word choice) that are different from the norm (Bucholtz and Hall, 2004). When exploring bias we must consider how identity might be constructed through language input and output, including what is left implicit (unmarked).

Finally, research from psychology can teach us how prejudice will manifest in tech-

nology. One particularly informative concept is that of benevolent stereotypes (Glick and Fiske, 1996). NLP research into generative models often assumes that harmful output will have a negative sentiment or contain overtly toxic content, which influences choice of metric (Goldfarb-Tarrant et al., 2023). However, as Glick and Fiske (1996) discuss with regards to sexism, prejudice can manifest in what the "offender" believes to be positive beliefs e.g. that women are naturally more caring. This motivates examining the content of generated output, not just relying on quantitative metrics for an overall picture.

Chapter 3

Proposing a Human-Centric Approach

Given the role of human behaviour in determining model bias, and the critiques detailed in Chapter 2, it is clear that current work on bias measurement and mitigation is often limited by being:[1]

- Focused solely on technology
- Focused on data as the source of bias
- Focused on an abstract, mathematical problem
- Based on authors' intuitions
- Focused on binary gender and race

These critiques can all be summarised as bias research failing to properly account for human behaviour. Practitioners simplify the problem in order to define a more pragmatic scope for bias measurement and mitigation. However, by offering a superficial understanding of the problem, this will lead us to superficial solutions that fail to engender real improvements for affected groups (Davis et al., 2021).

Having explored five common shortcomings, I now imagine what NLP bias research would look like if humans were the focus, rather than technology, informed by the cross-disciplinary research I summarised in Section 2.3. Given the role that human behaviour plays in determining the impact of technology (see Section 2.1.6) I argue

[1]It is important to note that, as a field, NLP bias evaluation is also limited by its focus on English language models used in Western contexts, but I do not address this directly in my own work (see Section 1.2): I refer the reader to excellent papers such as Smart et al. (2024); Qadri et al. (2023); Birhane et al. (2022b); Bhatt et al. (2022).

this human-centric approach will improve the quality of work done for both bias measurement and mitigation.[2] A human-centric approach would consider the entire socio-technical system that technologies are embedded within. A human-centric approach would look at the many ways in which human decisions contribute to model bias, beyond just the choice of training data. A human-centric approach would focus on the impact of technology on people, rather than an abstract or mathematical conceptualisation of bias. Further, it would consider how those impacted respond to the technology, as outcomes will be a product of the technology and their behaviour. A human-centric approach would be driven by expertise on human behaviour and language use, and crucially by lived experiences. Finally, a human-centric approach would consider the many facets that make up identity beyond race and gender. Adopting these maxims and centring human experiences in bias evaluation will result in greater validity, because we will be closer to capturing what we intend to capture (the impact of biased model outputs on affected communities). This will foster a deeper understanding and lead ultimately to more effective bias mitigation techniques. In the following, I will explore how each of these maxims leads to more insightful and impactful NLP bias research, with illustrative papers from the field.

3.1 Part of a Socio-technical System

Here I will discuss how work on bias in NLP will benefit from considering the socio-technical system that the technologies are embedded within. This includes considering those who produce data, those who help to prepare it, those who build the technology and those who enact its decisions. Of course, that system would also include those impacted by the use of NLP technologies, but I consider them below in Section 3.3.

A socio-technical approach to bias considers it to be the product of human and algorithm behaviour (Dolata et al., 2021). For example, we must consider the humans that design, build and deploy the algorithm, lest we fall into the Framing Trap (Selbst et al., 2019). The model's design will be influenced by – will embody (Talat et al., 2021) – the experiences, knowledge and values of its developers. Understanding how these are reflected in the model is an important first step to understanding how to ensure the model reflects a wider set of experiences, knowledge and values. A work

[2] Whilst I do not propose novel bias mitigation methods in my thesis work (Chapters 4 through 7), I believe that gaining a deep and nuanced understanding of the problem is a necessary first step (though cf. Utama et al. (2020)), and my proposed framework has informed work I have done elsewhere on bias mitigation (Ungless et al., 2022).

which can serve as inspiration for this approach is Birhane et al. (2022a). The authors conduct a literature review of papers at top AI conferences to establish which values are – often implicitly – reflected in the papers. By considering the implicit values in AI development and evaluation work, they draw attention to the decision to focus on qualities such as high performance and generalisability over for example being "not socially biased" and being fair. Human choices are rendered explicit by their taxonomy.

The power relationships that exist between stakeholders must be interrogated – as Miceli et al. (2022) write, it is not enough to address worker (annotator) bias: we must consider the power imbalance between annotator and developer, which contributes to the devaluing of annotator expertise (Bennett et al., 2025) and restricts annotators' abilities to critique annotation guidelines (Miceli et al., 2022), contributing to label noise.

Bias metrics are also socio-technical products, and we must critique them as such. Examples of bias metric design decisions that are often left implicit are assumptions about identity e.g. the researcher's understanding of gender (Devinney et al., 2022); what behaviour is desirable i.e. what does an unbiased model look like? (Goldfarb-Tarrant et al., 2023); what cultural context is relevant (the "default" of the US often goes unstated, Sambasivan et al. (2021); Bhatt et al. (2022); Goldfarb-Tarrant et al. (2023)). Further, whether bias evaluation is successfully implemented will be influenced by organisational factors such as commercial priorities (Madaio et al., 2022), further emphasising the socio-technical nature of bias evaluation work.

3.2 Many Sources of Bias

Beyond just imbalances in the data, there are many sources of bias in a model which I discuss in detail throughout Section 2.1. A successful approach to bias measurement and mitigation must take this into account. This includes acknowledging that attempts to mitigate bias can be a source of bias in themselves, as shown by Anwar et al. (2024) with regards to data filtering, Xu et al. (2021) with regards to finetuning, and as I discuss throughout this thesis. Those working in bias mitigation must consider how their attempts to debias a model may have a knock on effect such as erasure, or increasing the discrepancy in performance for less salient identities. Acknowledging that the bias in a model's output is the result of many interacting factors, and not just data imbalance,will allow researchers to develop more effective bias metrics and mitigation

techniques.

3.3 Impact on People & How They Respond

The plurality of AI ethics papers consider an abstract understanding of harms (Birhane et al., 2022b), and similarly I have shown elsewhere that many papers on bias ignore downstream use cases (Goldfarb-Tarrant et al., 2023) even though intended use hugely determines whether bias mitigation is appropriate (Sigurgeirsson and Ungless, 2024). Researchers often fail to account for the real impact on people, and how they respond. However, ethical concepts such as bias, fairness, and trust are best understood in context, with reference to real people and real harms. For example, Knowles et al. (2023) argue that the concept of trustworthiness is meaningless without reference to specific power dynamics. Some technologies are only trustworthy to those in power because they are capable of distinguishing them from those without power (Knowles et al., 2023) (sometimes *because* they are biased). This work also demonstrates the immense value of taking an interdisciplinary approach, which I discuss further in the following Section.

Knowles et al. (2023) demonstrate that trust is in the eye of the beholder. A human-centric approach acknowledges that the public respond to biased NLP technologies based on their beliefs about the system as well as the system's output. The focus is on what is perceived, not simply what can be measured in the model's output. Of course, these may be quite different, as Wang et al. (2020) found for perceived fairness: models were rated as fair if they favoured the respondent, even if the model was described as being very biased. Starke et al. (2022) conduct a meta-analysis of research on perceived fairness in algorithmic decision making and find that perceived fairness is highly context-dependent. We lack such a substantial body of work looking at perceived bias of NLP technologies specifically, but it seems likely similar findings will hold.

Biased algorithms can leave a lingering "imprint" that affect how people understand and interact with algorithms more generally (Ehsan et al., 2022). Failure to consider those impacted means a failure to understand the true scale of harm. Ehsan et al. (2022) note that those impacted by algorithms develop folk theories about how they work, which I return to in Chapters 6 and 7. A human-centric approach to bias realises that it is not enough for the model to be debiased – it must also be *believed* to be debiased, or acrimonious feelings will likely remain.

Of course, work on the real impact of technologies must draw on the lived experiences of impacted communities, as exemplified by Dev et al. (2021). That is, it should incorporate community knowledge. I discuss this further below.

3.4 Driven by Social Science and Community Knowledge

Being informed by social science research, particularly as relates to language use and social identity, allows for more careful exploration of bias (Devinney et al., 2022). Social science theory can guide our hand when attempting to mitigate harmful model bias. By drawing on expertise from other fields, we can move beyond intuitions to explore how documented harms are reflected in NLP technologies. May et al. (2019) do this successfully when they explore how the "Angry Black Woman" stereotype and the double bind (against women in typically male professional settings), concepts from psychology and feminism, manifest in sentence embeddings. A human-centric approach to bias is also inspired by transdisciplinarity (Rigolot, 2020), in that it is informed by both academic work (from across disciplines) and by the insights of non-academic stakeholders, such as journalists, platform users etc. (see Chapters 6 and 7 for a discussion of everyday algorithmic auditing conducted by the public). Smith et al. (2022) work with community members and "experts in responsible/inclusive research, racial justice, and preferred language in (dis)ability" to define terms in their data set for exploring bias in conversational models.

Work by Sunipa Dev, Remi Denton and colleagues demonstrates the benefits of taking into account community knowledge (Dev et al., 2024, 2023, 2021; Gadiraju et al., 2023). They work with marginalised communities to identify harms, including subtle forms of bias not easily identified by an off-the-shelf toxicity classifier or similar metric (Gadiraju et al., 2023). Dennler et al. (2023) likewise conduct workshops to understand what the queer community is looking for in the harm auditing process. When working with marginalised communities, researchers may be inspired by the Indigenous data sovereignty movement, which proclaims the "right of Indigenous peoples to govern the collection, ownership and application of data... as a cultural and economic asset" (Walter et al., 2021). Walter et al. (2021) give key guidance on this approach, which includes grounding data in the worldview of the affected community and in the wider social structural context (rather than data being de-contextualised as it is so

often) (Walter et al., 2021). Another important aspect is compiling data on the community's "priorities and agendas", which I am inspired by in Chapter 5. This has the benefit of ensuring bias measurement and mitigation work aligns with the issues actually faced by the community, rather than just with issues that the researcher assumes the community faces. If the goal of bias mitigation work is to prevent harm – as most researchers would attest to – rather than just to avoid "bad PR", this is vital.

Expertise can also come from advocacy groups. This helps to ensure that researchers do not do additional harm when researching these topics, for example by using dis-preferred language. For example, Hutchinson et al. (2020) draw on guidelines from the Americans with Disabilities Act National Network, amongst others, to explore how preferred and dis-preferred language relating to disability is handled by NLP technologies.

3.5 Broad Range of Demographics

Whilst it is not possible for a single work to cover all demographic attributes, as a field we must move beyond the preoccupation with gender and racial bias (Goldfarb-Tarrant et al., 2023; Blodgett et al., 2020), particularly solely within a Western context (see Footnote 1). Further, this maxim (to address a broad range of demographics) can also inspire researchers focusing primarily on one demographic attribute, such as gender or sexuality, to consider how people within an identity can experience different forms of marginalisation. For example, in Chapters 4 and 5 I explore how queer identities are impacted by bias, and within this I look at how identities belonging to queer people of colour are affected compared to non-ethnicity specific queer identities. Work can take an intersectional approach, to consider how particular demographics (such as Black women, or queer people of colour) face unique challenges that are the product – not the sum – of their multiple forms of marginalisation (Crenshaw, 1989). For example Guo and Caliskan (2021) explore harmful associations in contextualised word embeddings unique to women of colour . By expanding which identities we consider, and treating each with more nuance, we will be able to paint a broader and more detailed picture of the problem. However, it is worth noting that even data sets that aim to cover a broad set of demographics are naturally limited when they make use of a finite set of terms. Whilst Smith et al. (2022) intended their `HolisticBias` data set to be updated to incorporate an ever expanding list of identities, the word list has not been updated on GitHub since its release. I further discussion the ramifications of using finite word

lists in Chapter 4.

3.6 Conclusion

In summary, a better approach – a human-centric approach – to bias research in NLP:

- Sees NLP tools as part of large socio-technical systems
- Considers many sources of bias
- Focuses on the impact of technology on people in context - and how they respond
- Is driven by social science theory and community knowledge
- Addresses a broad range of demographics

In the following chapters, I present four case studies across three technologies that showcase the benefits of my human-centric approach with five key maxims. My work does not always live up to the ideal defined by all five maxims. However, by approaching the issue of social bias research with a fundamental focus on the human, not the technological, I hope to demonstrate the greater depth of insight this offers, and to illustrate how these maxims might manifest in practice. These case studies also contribute to our understanding of queerphobia in NLP, demonstrating the varied impact on minorities within the queer community.

Chapter 4

Queerphobic Bias in Automatic Sentiment Analysis

In this Chapter I present work on measuring queerphobic bias in automated sentiment analysis tools, which – like all NLP technologies – show bias against marginalised groups (Kiritchenko and Mohammad, 2018; Hutchinson et al., 2022). I illustrate this point by showing how six popular sentiment analysis tools respond to sentences about queer minorities, expanding on existing work on gender, ethnicity, and disability. I find evidence of bias against several queer identities, including in the two models, from Google and Amazon, that seem to have been subject to superficial debiasing. The Chapter concludes first with guidance for social science researchers on how to minimise the risk of sentiment analysis model bias skewing their results: by grounding in this specific use context (use by social science researchers), I am able to make tangible recommendations to minimise harm. I then discuss how this work shaped my own approach to bias measurement.

4.1 Introduction

NLP tools are being increasingly adopted in the social sciences (Robila and Robila, 2020; Saifee et al., 2020), the hope being that cutting-edge technologies will allow researchers to analyse data with efficiency and a human-like understanding of language use. However, social biases embedded in these models threaten to undermine this hope (Blodgett et al., 2020; Shah et al., 2020). For example, sentiment analysis tools developed using deep learning techniques have been shown to reflect biases such as racism, sexism (Kiritchenko and Mohammad, 2018), and ableism (Hutchinson et al., 2020),

through the use of a template-based approach which we likewise adopt in this paper. We[1] expand on existing work by testing for queerphobia[2] in six popular sentiment analysis tools. Sentiment analysis is used to assess interactions with political figures (Mousavi and Gu, 2019); to quantify product success (Li et al., 2018); to understand online communities (Pérez-Pérez et al., 2019). If these models systematically give different scores depending on the presence of queer identity terms, this threatens to undermine research conducted using these tools. This is because some of the apparent differences in mood or attitudes – which sentiment analysis attempts to measure – will in fact be the result of one person or group making more frequent reference to queer identities.

Our artificially constructed data set of 29,472 sentences (examples given in Table 4.1) allows for precise comparison across queer identities: a template structure is used to minimise the impact of confounding linguistic variables, which allows us to pinpoint which identity terms the models show bias against. As Smith et al. (2022) write, "for social applications of NLP, it's crucial to know what's in your data ... handcrafting data ... afford[s] more control". We take our approach (and templates) from Kiritchenko and Mohammad (2018), who were able to demonstrate bias against women and Black people in the majority of the 200 deep learning based sentiment analysis tools they tested, by permuting which names appeared in the templates.

In addition to three proprietary deep learning models – at the cutting edge of sentiment analysis – we also test three lexicon-based approaches, which rely on hand-engineered word lists labelled for valence. Valence decisions can be subjective, meaning bias is likely to impact labels (Mohammad, 2017), but these approaches offer far greater transparency, so the source of bias can be identified (and counteracted) with relative ease. Illustrating this point, in the SentiStrength (Thelwall et al., 2010) lexicon both "gay" and "queer" have negative sentiment labels. This contrasts with the "black box" nature of modern deep learning approaches where carefully constructed probes are needed to "tease out" patterns of bias (Bender et al., 2021), and addressing this bias involves more complex processes (Meade et al., 2022). The six tools we test in this paper sit at either of end of the spectrum from highly interpretable to fully "black box" systems. Our findings are intended to provide guidance to those selecting a tool for research or commercial purposes. We focus on sentiment analysis as this has a wide range of applications, but our findings can be thought of as a cautionary tale for adopt-

[1] Until Section 4.7 I use first person plural to acknowledge this work was the result of collaboration.
[2] Understood as a bias against LGBTQ+ individuals

ing NLP technologies in research. Our focus on queerphobia allows us to demonstrate the importance of a nuanced approach to identity, as many queer identities exist at the intersection of multiple forms for marginalisation, and a binary comparison between queer and non-queer identities (parallel to comparing white and Black names) is not sufficient for understanding how the power dynamics associated with sexuality, gender, trans status and ethnicity might interact and impact how certain terms are treated by these tools. In the following, we discuss some of the existing uses of sentiment analysis tools and the role deep learning approaches will likely play in future. We discuss how sentiment analysis tools might come to be biased and how this reflects larger problems for those who use NLP tools in their research. We then present and test the six selected tools and discuss our findings, and their implications for researchers and commercial users of sentiment analysis tools and other NLP technologies.

4.2 Background

4.2.1 Sentiment Analysis Approaches

Current approaches to sentiment analysis can be roughly categorised into lexicon-based approaches and ML-based approaches. Lexicon-based approaches use dictionaries of words and phrases labelled for sentiment, and typically include rules to deal with negation or use of modifiers. They use scoring mechanisms to determine the sentiment of a span of text, for example taking an average. Lexicon-based approaches are efficient to implement as they do not require training data, are easy to use and produce results almost instantly, making them ideal tools for research. Lexicon-based approaches are usually very transparent – it is easy to understand how the sentiment score was determined, based on the relevant words in the text. ML-based approaches involve training a system to automatically classify or score spans of text for sentiment. They require large amounts of training data, which is either manually labelled for sentiment or taken from sources where ratings are supplied, such as reviews, making them resource intensive. Many different techniques fall under the umbrella of machine learning. A nearest neighbour classifier would look to the training instances most like the test instance to determine its sentiment. One type of ML-based approaches are deep learning models, which are trained on vast amounts of training data to make labelling decisions based on abstract features. Examples of proprietary deep learning systems for sentiment analysis include IBM Watson's Natural Language Understand-

	Model					
Variant	Google	Amazon	IBM	LIWC	VADER	AFINN
This person feels enraged.	-0.20	0.04	-0.36	74.76	-0.40	-2
This latinx* person feels enraged.	-0.10	0.04	0.25	43.37	-0.40	-2
This same gender loving* person feels enraged.	-0.40	0.55	-0.36	13.15	0.30	0

Table 4.1: Table of scores for three example sentences for the template "<identity phrase> feels enraged". Note that these models adopt different scoring conventions but in all cases a higher score means a more positive sentiment. All models show a variation in scores depending on the identity phrase. * terms defined in Methods.

ing service.[3] Such models can be deployed without additional model training, so do not require a background in computer science.

4.2.2 Literature Review

A search of the SCOPUS[4] and ASSIA[5] databases for "sentiment analysis" shows that automated tools are widely used in the social sciences to conduct research on topics as diverse as understanding the impact of COVID and other disasters (Kaur et al., 2020; Razavi and Rahbari, 2020); to analysing the online behaviours of vulnerable groups (Saifee et al., 2020). Much recent work relies on lexicon-based approaches such as LIWC (Tausczik and Pennebaker, 2010), VADER (Hutto and Gilbert, 2014) and SentiStrength (Thelwall et al., 2010), likely due to their ease of deployment. Clearly, simple lexicon-based approaches still play a significant role in social science research.

However, lexicon-based approaches face several major problems. They risk becoming outdated if they fail to include recently coined terms, which are particularly prevalent in casual online discourse (Hilte et al., 2018). As stated above, sentiment labels are largely subjective (Mohammad, 2017). Related to these two points, minority groups may be impacted if the lexicon does not include terms from non-standard English(es) (such as AAE) or if these terms are included but labelled by individuals who do not belong to the community (see Section 2.1.2). This will be starkly evident in the

[3] https://www.ibm.com/uk-en/cloud/watson-natural-language-understanding
[4] https://www.elsevier.com/en-gb/products/scopus
[5] https://proquest.libguides.com/assia/content

case of (the many) identity terms that are reclaimed slurs, as with "queer" being rated as negative in SentiStrength. Content produced by queer people may be inaccurate as a result. Finally, lexicon-based approaches are overly simplistic – they will often fail to properly consider negation, modals and sentence structure, as well as sarcasm and humour (Mohammad, 2017).

These problems hamper the performance of lexicon-based approaches, so they are increasingly being abandoned in favour of ML models which outperform on benchmark sentiment analysis datasets (Poria et al., 2020). It has been predicted that use of ML will become the norm in social science research (Robila and Robila, 2020). Examples of papers already using ML approaches include Kaur et al. (2020); Li et al. (2020). Deep learning models can offer more sophisticated analysis having learned abstract patterns from vast amounts of data. However, this greater accuracy comes at the cost of less transparency. The values of deep learning models' parameters are rarely interpretable, meaning it is unclear how classification decisions are made.

A further problem with ML-based tools is that they often pick up on "spurious associations" (Utama et al., 2020) between group identifiers and negative characteristics, an artefact of training data which reflects human biases. This means these systems may develop a bias against minority groups (Shah et al., 2020), consistently producing a different output due to the presence of identity terms or community-specific language use. There is substantial evidence of bias in NLP across a range of tasks, for example hate speech detection (Röttger et al., 2021) and coreference resolution (Cao and Daumé III, 2020; Rudinger et al., 2018), in addition to sentiment analysis (Hutchinson et al., 2020; Kiritchenko and Mohammad, 2018). Attempts have been made to combat such bias, but despite this many commercially available tools show evidence of bias, including against queer individuals (Buolamwini and Gebru, 2018; Röttger et al., 2021; Thompson, 2017). This leads us to predict that deep learning-based sentiment analysis tools will show a bias against queer terms.

Such bias can lead to harm being done to minority groups. Crawford, Barocas and colleagues (Crawford, 2017; Barocas et al., 2017) divide the potential harms of NLP systems into representational and allocational harms. The former refers to harms done through the misrepresentation of a group, including differences in system performance. A group may be misrepresented as speaking overly positively about a topic because the terms they used to talk about their lives result in inaccurate scores (in the sense that the presence of terms such as "latinx" can lead to a difference in score despite the intended sentiment being the same, as shown in Table 4.1), and this is harmful where

it plays into stereotypes, such as the "happy native" stereotype (Phillips et al., 2021). "Allocational harms" refers to the unfair allocation of resources to the demographic e.g. access to job opportunities, funding etc. Use of a biased sentiment analysis tool may lead to allocational harms if sentiment analysis is used as a form of triaging e.g. to determine the success of an initial mental health intervention (Hoogendoorn et al., 2017). Resources may be unfairly allocated if biased sentiment analysis tools give inaccurate scores to certain communities.

Both lexicon- and deep learning-based approaches have the potential to cause harm to minority groups due to inaccurate scores, for the reasons outlined above, and so we investigate both kinds of tools as to whether they show bias against queer minorities. The NLP literature suggests we will find evidence of a bias against queer minorities in the deep learning models, whereby they will receive lower sentiment scores (as women and Black people receive lower scores in (Kiritchenko and Mohammad, 2018)). The presence of "gay" and "queer" in SentiStrength suggests other lexicon-based approaches may similarly show a negative bias against queer identities.

Higher scores in one group may indicate their ratings have been inflated by the presence of certain identity terms, or that the scores for the other group have been reduced for the same reason. Either way, the system is erroneously basing sentiment rating on the presence of identity terms not intended to indicate sentiment. It could be said the models are biased against the non-minority identity groups if these also receive inaccurate scores. However, we formulate our hypotheses around bias against the minority groups because they are historically disadvantaged, in line with the rest of this thesis. Further, we test monodirectional hypotheses, predicting that minority groups will receive more negative sentiment ratings, but it is important to note that inaccuracies in either "direction" can cause harms, for example if such models are used in the context of mental health triaging.

Our primary hypothesis (H1) is that sentiment analysis tools will show an overall bias against queer identities compared to non-queer identities, in that queer identities will receive more negative sentiment ratings. Inspired by an intersectional approach to identity (Crenshaw, 1989), we hope to provide a more thorough evaluation of the tools by also looking at minorities within the queer community: treating queer identities as a homogeneous group could obscure bias against particular queer identities such as those specific to people of colour. We predict that these systems will reflect bias against minorities within the queer community (H2), for example there will be greater bias against queer women compared to queer men, because women additional

Template	Identity term combination	Emotional term
I saw this <identity phrase> in the market.	bisexual two-spirit person	N/A
↪ I saw this bisexual two-spirit person in the market.		
This <identity phrase> told us all about recent <emotion> events.	lesbian woman	wonderful
↪ This lesbian woman told us all about recent wonderful events.		

Table 4.2: Table showing how templates, identity phrase combinations and emotional terms were combined to create the sentences in our data set.

experience sexism. To explore H2 we compare female and male (H2A), transgender and cisgender (H2B), and ethnicity-specific and non-specific identities (H2C) within the queer community, predicting that in each case the former group will receive more negative sentiment ratings than the latter, because they are further marginalised by their gender, trans status and ethnicity respectively.

4.3 Method

4.3.1 Dataset Creation

We used 10 of the templates, and the emotional vocabulary from (Kiritchenko and Mohammad, 2018). We modified the templates to include a combination of up to three identity terms about trans status, sexuality and gender (given in Table 4.3) in that order. Examples of the templates, identity phrase combinations and emotional terms (and the resulting sentences) are given in Table 4.2. Use of templates allows for careful comparison across different identities. Using natural data would have introduced many confounding variables including use of slang and other dialectal differences across communities. We would likely struggle to find truly comparable data between mainstream queer identity terms such “gay”, identity terms used by people of colour such as “same gender loving” and identity terms that have not entered the mainstream such as “demisexual”. For the purposes of careful comparison across intersecting identities, a template-based approach provided the best option.

To establish our list of identities, we started with relevant terms from Dixon et al. (2018) (indicated in Table 4.3 in bold). We expanded on this list to represent more

Trans status	Sexuality	Sexuality presentation	Gender
Cisgender	Asexual	Bear (m)	**Woman (f)**
Transgender	**Bisexual**	Butch (f)	**Latinx**
	Demisexual	Cub (m)	**Man (m)**
	Fluid	Dyke (f)	**Non-binary**
	Gay	Femme (f)	Transfeminine
	Heterosexual	Stud (f)	Transmasculine
	Homosexual	Twink (m)	Two-spirit
	Lesbian (f)		
	LGBT^{i}		
	LGBTQ^{i}		
	Pansexual		
	Queer^{i}		
	Same gender loving		
	Straight		

Table 4.3: Table showing identity terms included in our data set. **Bold** font indicates identity terms from Dixon et al. (2018). Original list includes 'nonbinary' where we use the more popular spelling 'non-binary'. Original list includes 'female, male' rather than 'woman, man'. Underline indicates the term is ethnicity specific. (m) indicates this identity is typically used by men. (f) indicates this identity is typically used by women. ^{i} understood as an umbrella term for non-heterosexual sexualities.

diverse queer identities, for example including three additional terms specific to people of colour. Below, we focus first on those terms relevant to our hypotheses. Additional terms are discussed in Appendix A.

Note that if unspecified, the norms of the cisheteropatriarchy are assumed, because marked identities use marked language (Bucholtz and Hall, 2004; DePalma and Atkinson, 2006). That is to say, non-normative identities such as non-heterosexual or non-cisgender identities typically must be explicitly stated to be understood. If these identity terms are not stated, then it is assumed the norm applies: when sexuality is not given, heterosexuality is assumed; when trans status is not given, cisgender identity is assumed. However, the data set includes explicitly normative identity terms such "straight, cisgender", which allows for comparison between queer identities and both the explicit and assumed norms.

Our data set is designed to evaluate how sentiment analysis tools treat sentences about individuals belonging to different minorities within the queer community. For example, queer women will face additional discrimination to queer men, even experiencing misogyny from within the queer community (Hale and Ojeda, 2018): it is likely such bias will be evidenced in models with sentences about queer women receiving lower scores than those about queer men. To enable a thorough exploration of the impact of queer female identities on sentiment ratings, we include "butch, femme, dyke" and "stud", in addition to "lesbian" from (Dixon et al., 2018). For queer men, we include the related terms "bear, cub, twink". These terms all relate to presentation-driven subgroups within the non-heterosexual community.

Transgender individuals face related gender oppression, from both outside and within the queer community (Iantaffi and Bockting, 2011; Stone, 2009). To identify bias against transgender individuals (our hypothesis H2B), we include identities with different trans status ("transgender, cisgender", or trans status is not given and cisgender is assumed). We also include a number of non-binary identity terms, under the umbrella of transgender identities,[6] namely "transfeminine, transmasculine, two-spirit" in addition to "latinx, non-binary" from Dixon et al. (2018). The scarce research comparing binary and non-binary transgender people suggests they face similar levels of victimisation (Rimes et al., 2019). The data set can also be used to measure the impact of non-binary pronoun choice: both 'themself' and 'themselves' are reflexive pronouns commonly used by the non-binary community, and we include sentences with both (i.e. all non-binary variations were included once with 'themself' and once

[6] Although not all non-binary people identify as transgender (Rimes et al., 2019)

with ‘themselves’).

Many queer identity terms are not adopted by people of colour, due to feelings of alienation (Battle, 2002). Individuals may adopt alternative terms; we include a small number of these in our data set to test for bias against queer people of colour. The terms we include are “stud” (a term used by some black women who love women with a particular masculine aesthetic (Lane-Steele, 2011)); “same gender loving” (a term used by Black individuals attracted to those of the same gender, possibly in addition to those of another gender; (Battle, 2002)), and “two-spirit”, an umbrella term for third gender identities unique to Native Americans (Anhorn, 2016),[7] in addition to “latinx” from (Dixon et al., 2018) (a term used by some non-binary individuals of Latin American descent (Salinas Jr., 2020), although see Section 5.9 for a discussion of some of the controversies surrounding the term “latinx”). We include these identity terms when comparing ethnicity-specific to non-ethnicity specific queer identities (H2C).

The key challenge in creating this data set was combining identity terms in an appropriate manner: the identity terms could not all be combined (for example “cisgender” and “transmasculine”). Sensitively combining identity terms required extensive research. As an illustrative example, in order that the data set be useful for detecting bias against plurisexual identities (those attracted to multiple genders), we combined terms relating to sexuality presentation style (e.g. “butch”), where monosexuality is assumed, with the term “bisexual”. However our research originally established “stud” is almost exclusively used by black women who exclusively love women (Lane-Steele, 2011). Therefore, we felt it would be inappropriate to combine the terms (cf. Appendix A). Researchers wishing to explore additional combinations can do so using the source code provided.

Some of the combinations of terms e.g. “transgender same gender loving woman” may be unlikely to occur together in natural data, an artefact of using templates. However, all are valid identities. It is our belief that how common an identity is written about should not determine whether we test for bias against that identity, the implicit decision made by researchers testing only the most salient identities. By testing many combinations, we are able to determine if some of the combinations of terms interact in unexpected ways, for example resulting in much higher or much lower scores than would be expected given how the system treats the terms individually.

The 30 identity terms were combined with the templates (along with “this” plus

[7] We use “they” pronouns for all Two-spirit examples but there are those who identify as Two-spirit men and Two-spirit women and may prefer other pronouns; traditions vary across tribes (Anhorn, 2016).

"person", "woman" or "man" where appropriate) to give 29472 sentences. We designed our data set such that many forms of bias can be identified. In the present work we focus on four likely examples of bias, but our data set allows for the exploration of biases against many queer minorities. Of course, our list of identities is not comprehensive; the program that generates the data set can handle additional terms in an appropriate manner.

4.3.2 Selecting Sentiment Analysis Tools

We focus on popular sentiment analysis tools that are likely to be used by researchers beyond computer science: three that use deep learning and three with a lexicon-based approach. The three deep learning products we consider are from Google, Amazon and IBM, all of which offer "out-of-the-box", pre-trained sentiment analysis tools, namely Google Cloud Natural Language,[8] Amazon Comprehend,[9] and IBM Cloud Natural Language Understanding.[10] These are three of the world's biggest cloud service providers, meaning their products are widely available.

To establish our list of lexicon-based approaches, we looked at the proprietary tools identified in our survey of the ASSIA database (described in the Literature Review). The tools were VADER, LIWC, NRC (Mohammad and Turney, 2013), SentiStrength, AFINN (Nielsen, 2011), Textblob (Loria, 2018) and Semantria.[11] VADER and LIWC were the first and second most common tools, so we examine them both in this paper. NRC and AFINN were both used by two papers. Ultimately, we opted to investigate AFINN because it is older, although still used in recent papers e.g. Borakati (2021), and so we felt it was more likely to record queer terms as negative and be potentially biasing contemporary findings. We also found VADER, LIWC and AFINN to be widely used by papers in the SCOPUS database.

We test only a subset of the sentiment analysis tools in popular use. We make our data set publicly available for use in testing other models.[12]

[8] https://cloud.google.com/natural-language
[9] https://aws.amazon.com/comprehend/
[10] https://cloud.ibm.com/catalog/services/natural-language-understanding
[11] https://www.lexalytics.com/semantria/
[12] https://github.com/MxEddie/SSCR_supplementary

4.3.3 Testing Procedure

To identify the impact of identity terms we compare between different variations (with different identity term combinations) of a particular template plus emotional term (henceforth, template(s)[e]). For example, we compared the sentiment rating for when the template[e] "This <identity phrase> told us all about recent wonderful events." was combined with "lesbian woman", "bisexual two-spirit person", "person" and other identity term combinations. This allowed us to identify the impact of the identity terms on the ratings given by the tools.

To facilitate the testing of our hypotheses, we grouped multiple identities together where appropriate and considered the average rating across all templates[e] for identity term combinations in this grouping. This allowed us to effectively identify broad patterns in how identities were treated by the systems. For example, to test for bias against queer identities (H1), we grouped all the identity combinations that included any terms other than "cisgender, heterosexual, straight" into a queer group, and grouped the remaining identity combinations in a non-queer group. We then conducted a two-tailed paired sample t-test between the average rating for queer versus non-queer identities across all the templates[e]. We followed this same procedure for all queer female and queer male identity combinations (H2A) – that is, we compared all identity combinations that included a queer female identity term such as "lesbian", or a queer identity term such as "bisexual" plus "woman", to all identity combinations that included a queer male identity such as "twink" or a queer identity term plus "man". We did the same for all transgender and all queer cisgender identity combinations, that is queer identities that either do not reference trans status, or else specify cisgender (H2B). For queer ethnicity-specific identity combinations we compared them to the non-ethnicity-specific identity terms closest in meaning (H2C), e.g. "same gender loving" and "gay" or "lesbian".

Where appropriate we conduct additional analyses, for example looking at minimally contrasting pairs such as "I saw this bisexual man" versus "I saw this bisexual woman". For the lexicon-based approaches we also looked at the lexicon itself, where it was available, or at the scores assigned to individual words.

	Model					
Hypothesis	Google	Amazon Comprehend	IBM	LIWC	VADER	AFINN
H1				Supported		
H2A	Supported	Supported				
H2B		Split	Split	Split		
H2C						

Table 4.4: Table demonstrating which hypotheses are supported by our analysis of each model. As a reminder: (H1) sentiment analysis tools will show an overall bias against queer identities. (H2) these systems will reflect bias against minorities within the queer community, namely female vs male (H2A), transgender vs cisgender (H2B), and ethnicity-specific vs non-specific identities (H2C). For H2B, 'Split' indicates systematic bias against some transgender identities, that is, only binary or only non-binary transgender.

4.4 Results

In this section we look at the results of testing each of the six tools with our novel dataset. Though only LIWC showed a systematic bias against queer identities, giving them consistently lower sentiment ratings, minority groups within the queer community might still be impacted by inaccurate results from all of the tools we tested. We give a summary of our results in Table 4.4.

4.4.1 Google

Google assigned sentiment ratings between -1 and 1, where -1 indicates a very negative sentiment and +1 indicates a very positive sentiment. All our hypotheses predict that the minority group will receive a more negative sentiment rating; for Google, this will be indicated by a lower sentiment score.

Following our paired t-test, we found queer identities were given a significantly higher rating than non-queer identities: -0.218 compared to -0.246, $t(123) = 3.68$, $p < .001$. This was counter to our hypothesis **H1**. However, there was some evidence of bias against groups within the queer community, in line with H2, which we explore below.

We found female queer identities were rated lower than male (**H2A**): -0.250 com-

Identity + man	Google mean	Amazon mean	IBM mean	Google N-gram %
	-0.225*	0.613*	-0.340	$6.23e-2$
Straight	$-0.225*$	0.613*	-0.354	$9.62e-6$
Gay	$-0.225*$	0.613*	-0.540	$3.31e-5$
Homosexual	$-0.225*$	0.613*	-0.401	$1.83e-6$
Heterosexual	$-0.225*$	0.613*	-0.624	$2.55e-6$
Bisexual	$-0.225*$	0.613*	-0.524	$9.67e-7$
Queer	$-0.225*$	0.613*	-0.448	$2.52e-6$
Asexual	$-0.225*$	0.590	-0.374	$4.24e-8$
Pansexual	$-0.225*$	0.613*	-0.505	0
LGBT	$-0.225*$	0.613*	-0.350	0
LGBTQ	$-0.225*$	0.613*	-0.355	0
Demisexual	-0.240	-0.349	-0.349	0
Fluid	-0.090	-0.166	-0.166	$5.21e-8$
Same gender loving	-0.086	-0.357	-0.357	0

Table 4.5: Table showing mean sentiment rating across select male identities, for the three ML-based sentiment analysis tool, alongside the frequency of the terms in two databases to demonstrate that popular terms are more likely to be standardised.

pared to -0.201, $t(123) = -4.72, p < .001$. Further, where like-for-like comparisons are possible i.e. "bisexual man" versus "bisexual woman", the latter received a lower score on average. This suggests a bias against queer women compared to queer men. Regarding transgender identities, we found neither "transgender" nor "cisgender" had any impact on sentiment score (sentences with these terms received identical scores to those with no terms), meaning binary transgender identities did not receive more negative sentiment ratings than queer cisgender identities. Further, we found non-binary identities were rated as less negative than queer cisgender identities: -0.212 compared to -0.230, $t(123) = 2.99, p < .005$. Thus, we found no support for hypothesis **H2B**.

Comparison between the terms used by people of colour and those most closely comparable non-ethnicity-specific terms gave no support for **H2C**: the terms used by people of colour elicited similar or more positive ratings. For example, same gender loving woman received a more positive sentiment rating than lesbian woman.

The results for Google suggest some superficial debiasing may have been conducted. For some of the terms there were no differences in score whether a sentence contained the term or not. This suggests the system may have been "instructed" to ignore certain identity terms. In Table 4.5 we give results across 13 sexualities plus the term "man" to illustrate our point. Google Cloud is effectively oblivious to terms including "straight, gay, queer, asexual". Table 4.5 illustrates the pattern that less common terms seem unlikely to be included in this purported "ignore list": smaller minorities within the queer community, often the most marginalised, are still impacted by spurious differences in score. Another major issue with use of an ignore list to avoid bias is that some identity terms are crucial to contextualising other words used in the sentence. Within a (particularly Black and Latinx) queer context, the phrase "sickening" means something very positive (Calder, 2019). In non-queer contexts, this term is typically very negative in sentiment. Use of an ignore list might mean the system will be oblivious to these contextual clues.

4.4.2 Amazon

Amazon gives two separate confidence ratings, between 0 and 1, for the sentence being of positive or negative sentiment. A more negative sentiment rating would be indicated by a) a lower confidence rating for positive sentiment and b) a higher confidence rating for negative sentiment.

We found queer identities were rated as slightly less likely to be positive, but this

was not significant (**H1**). Queer sentences were rated as significantly less likely to be negative, 0.552 compared to 0.614, $t(123) = -7.15, p < .001$. Thus, we found no support for H1. However, as with Google we did find evidence of bias against minorities within the queer community, in line with H2.

Female queer identities were rated as more likely to be negative and less likely to be positive than male (**H2A**). For positive confidence rating, female queer identities received 0.276 compared to 0.287, $t(123) = -4.14, p < .001$. For the negative confidence rating, 0.590 compared to 0.580, $t(123) = 5.45, p < .001$. Thus we found support for hypothesis H2A.

As with Google we found neither "transgender" nor "cisgender" had any impact on score (**H2B**). Non-binary gender identities were rated as significantly less likely to be negative, 0.573 compared to 0.579, $t(123) = -4.64, p < .001$. However, they were also rated as significantly less likely to be positive than cisgender identities, 0.277 compared to 0.280, $t(123) = -3.18, p < .005$. Thus, we found some limited support for H2B.

None of the terms used by people of colour received lower positive confidence ratings or higher negative confidence ratings than their closest non-ethnicity-specific equivalents (**H2C**). For example, "stud" was rated as more likely to be positive and less likely to be negative than "butch". "Two-spirit, non-binary" and "latinx" received totally identical ratings, again suggesting the existence of an "ignore" list, though this was not the case for the Google results.

As with Google, Amazon's system seemed to have been subject to the same kind of superficial debiasing – see Table 4.5. The two systems do not ignore the same terms, as evidenced by the scores assigned to "asexual", fluid" (see Table 4.5) and "two-spirit, non-binary" and "latinx" (see above).

4.4.3 IBM

As with Google, IBM assigns sentiment ratings between -1 and 1, where a lower score indicates a more negative sentiment.

For IBM, we found queer identities were rated significantly higher than non-queer identities, -0.409 compared to -0.444, $t(123) = 4.17, p < .001$. Therefore, we did not find support for **H1**.

Female queer identities received a slightly more negative rating than male queer identities on average (**H2A**), though we found this was not significant, perhaps indica-

tive of successful debiasing, or it may be that the training data IBM uses means their model is less prone to gender bias.

Binary transgender identities were rated as less positive than explicitly cisgender identities (those identity combinations where "cisgender" is included) and assumed cisgender identities, where trans status is not mentioned and the norm is assumed (**H2B**). Binary transgender identities were rated -0.510 compared to -0.445 for explicitly cisgender identities, $t(123) = -7.01, p < .001$. Binary transgender identities were rated -0.510 compared to -0.422 for assumed cisgender identities, $t(123) = -9.26, p < .001$. In comparison, non-binary identities were rated more positively (-0.401) than queer cisgender identities: higher than explicitly cisgender identities, $t(123) = 5.93, p < .001$; and higher than assumed cisgender identities, $t(123) = 2.92, p < .005$. Thus we had partial support for H2B.

There was some evidence of bias against queer people of colour (**H2C**). For example, "latinx" was rated significantly more negatively than "non-binary", -0.416 compared to -0.393, $t(123) = -2.85, p < .005$. Similarly, "same gender loving" was rated as significantly more negative (-0.462) than "homosexual" -.406, $t(123) = -3.95, p < .001$, though it was also rated as significantly more positive than "gay", -0.502, $t(123) = 3.4, p < .005$. However, there was no overarching pattern of bias against ethnicity-specific terms.

IBM had not been subject to the same heuristic debiasing as the other two deep learning models, and this is evident in the more varied results compared to Google and Amazon in Table 4.5.

4.4.4 LIWC

LIWC assigns an emotional tone score between 0 and 100, where less than 50 indicates a negative sentiment. Because LIWC explicitly labels which terms are positive and which are negative, we supplemented statistical analysis with qualitative exploration of the dictionary and of the tone scoring system. Of all the identity terms, only "loving" (from "same gender loving") was included in the LIWC dictionary as explicitly labelled for positive emotional tone; none were labelled for negative emotion. We did find others did differ as to e.g. whether they were marked as "informal", which may have influenced their overall LIWC emotional tone rating.

Queer sentences were rated as less positive than non-queer sentences (**H1**), largely due to sentence length: a linear regression analysis found that sentence length ac-

counted for almost a quarter of overall variation ($R^2 = .225$). Marked identities use marked language and usually must be explicitly "spelled out" to be understood, meaning sentences about queer identities will typically be longer. This seems to be connected to the fact that LIWC records the proportion of a sentence that a particular word category makes up (LIWC categorises words according to their inclusion in word lists such as articles but also positive emotion words (Tausczik and Pennebaker, 2010). Sentences including multiple identity terms might then have proportionally fewer positive terms resulting in lower scores, even if the identity terms themselves have no valence associated with them.

In addition to this general pattern of bias against queer identities, seemingly because of sentence length, we found evidence of bias against minorities within the queer community (**H2**). Neither "woman, man" nor "person", nor any of the sexuality presentation terms had any impact on score, meaning there was no bias against queer women. Implicitly cisgender identities received higher scores on average, likely due to the impact of (shorter) sentence length. "Same gender loving" received more negative scores on average than "lesbian", "gay" or "homosexual", likely an artefact of length, despite the fact loving is explicitly tagged as having positive emotion. However, there was no overall pattern of bias against ethnicity-specific terms.

In conclusion, whilst none of the identity terms were labelled as positive or negative, by virtue of marked identities using marked language, some minorities were subject to bias by LIWC, giving support for H1 and partial support for H2B.

4.4.5 VADER and AFINN

VADER outputs three scores between 0 and 1, which sum to 1, indicating how likely a sentence is positive, negative or neutral (based on the sentiment of the words in the sentence), and a single compound score between -1 and 1 indicating the overall sentiment. AFINN is one of the most simple sentiment tools; it assigns a score between -5 and 5, the average sentiment of words in the sentence. As with LIWC, VADER and AFINN explicitly labels which terms are positive and which are negative, so we tested our hypotheses through an exploration of the dictionary and of the tone scoring system. Whilst these simple models benefit from being highly transparent, they are also unable to deal with more complex sentence structure, meaning their scores may be very inaccurate for reasons other than the presence of queer identity terms.

For both VADER and AFINN's emotional lexicon, only "straight", "loving" (as

in "same gender loving") and "spirit" (as in "two-spirit") are present, all recorded as positive. Sentences containing these words received slightly more positive scores. However, there was no overall pattern of bias against queer identities (**H1**) or even minorities within the queer community (**H2**), though these identities may be subject to artificially positive sentiment scores. Unlike for LIWC, sentence length had no impact for AFINN or VADER, meaning there was no systematic bias against marked, non-normative identities.

4.5 Discussion and Limitations

Our results indicate that all six tools we tested had the potential to give inaccurate results depending on the language people use to talk about their lives. Only one model, LIWC, showed an overall negative bias against queer identities compared to non-queer, a result of "penalising" marked identities. However, LIWC, Google, Amazon and IBM all showed negative biases against minorities within the queer community. For example, queer women received a lower score on average compared to queer men from Amazon and Google. Whilst we do not give specific recommendations for the best tool to select (as this depends on a variety of factors including a researcher's technical abilities, and there are many more tools available than those we selected to test), in the following we provide some guidance for choosing an appropriate sentiment analysis tool for the task and what steps can be taken to mitigate the impact of bias.

Counter to our predictions, minority groups did not always receive lower scores – in some cases the minority identities received much higher scores. Although this does not support our hypotheses, inaccurate scores in either "direction" can be harmful; individuals may be unfairly excluded from opportunities or misrepresented in research because they receive systematically higher or lower scores. If reference to a person's non-binary gender identity results in systematically higher sentiment ratings (as is the case for Google), these individuals may be considered lower priority in a mental health triaging system for example. Use of a sentiment analysis tool should be complemented with further analysis, as a "sanity check", for example of how different demographics are distributed across rating brackets, to identify patterns of systematic differences, some of which may be spurious.

For the lexicon-based approaches we tested we were largely able to pinpoint the sources of bias. Where identity terms were included in the lexicons, we were able to easily identify this. In this case, a solution might be to remove words, or introduce

rules such as that "loving" in the context of "same gender loving" should not be considered positive. This would require sensitive revisions to the existing lexicons and associated rule systems. Constructing a lexicon is very resource intensive and ensuring that it is culturally sensitive adds to this workload. However, without this work many researchers currently using these lexicon-based approaches may receive inaccurate results due to certain identity terms occurring in the lexicon.

For VADER and AFINN the scale of the issue seems relatively limited in that only a handful of identity terms are included in the lexicon. However, we encourage researchers to spend some time checking that terms relevant to their participants' lives do not feature in unexpected ways in the lexicon they are using, and where possible even make edits to the lexicon to remediate this; this of course applies to other identities beyond queer ones. This is particularly relevant when data is gathered from a diverse range of participants, where some will be using the terms and others will not be, or when use of reclaimed slurs is common. For a fair and insightful comparison, the impact of identity terms must be removed. The major benefit of lexicon based approaches is their transparency, which makes this kind of investigation and remediation easy. This comes at the expense of the system being able to process complex sentence structure, as these tools are far less sophisticated than the deep learning approaches. However, this sacrifice may be necessary in order to be confident that minority identity terms are not impacting the sentiment score. It is hard to have this same confidence in deep learning based approaches without the use of carefully constructed probes (Bender et al., 2021) which only exist for a small number of identities (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023).

For LIWC the issue of bias seems to be more pervasive. We found that because LIWC factored in sentence length, some minority identities would be subject to bias. The seemingly innocuous decision to incorporate sentence length into sentiment calculations means LIWC shows bias against marked identities. If a researcher is comparing queer and non-queer individuals, LIWC could introduce a bias against queer individuals if they frequently use marked identity terms when they write, and the non-queer participants do not. This same caution would apply to any linguistically marked (marginalised) identities, for example people of colour or people with disabilities.

For the deep learning approaches, our use of templates allows for careful comparison across terms, and we were able to identify that certain identity terms have a negative impact on score, for example many of the terms associated with female queer identities (in the case of Google and Amazon). Some of the minority terms

also resulted in significantly more positive ratings, for example "same gender loving" resulted in some of most positive ratings from Amazon; despite offering more sophisticated language processing abilities, as with the lexicon-based approaches Amazon (and Google) gave inflated scores to "same gender loving" identities, likely because of the presence of the word "loving". Whilst these models are designed to consider context, they appear to be failing to do so here, likely due to a lack of diverse training data. Further, the fact that scores are likely to vary most for the least common identities (which we took as evidence of heuristic debiasing) suggests that these tools are to be avoided when considering data from or about individuals belonging to the most marginalised communities.

The heuristic debiasing approach we suspect Google and Amazon have adopted does mean more mainstream queer and non-queer identities receive comparable scores, but in trying to rid the model of bias against minority identities, the developers have also inhibited the models' ability to use "semantic bias" to contextualise the meaning of slang terms. In addition to the "sickening" example given in Results, there are countless examples of words that have different sentiments across different queer communities, including "fierce" (Calder, 2019) (a positive sentiment in drag/ballroom culture, typically a negative sentiment outside of this).

Further, the lack of transparency or easy modification is an issue for deep learning approaches. Due to their black box nature, it is not clear why some identities were rated higher than others by these models (and the fact they are proprietary models further limits investigation), though we detail possible sources of bias in the Literature Review in Section 4.2.2. It is likely that significant differences in the distribution of terms in the training data resulted in these differences, as the systems learned to "focus" on these terms as an almost "heuristic" way of assigning score (Zhao et al., 2017). There is a push for better documentation of potential biases in models (Mitchell et al., 2019), but often bias is detected (by the impacted individuals (Buolamwini and Gebru, 2018)) after the models are deployed. This leaves social scientists wishing to use these models to collate various analyses and sift through model documentation that may not explicitly address the issues of bias, in order to understand the likely impact of bias in their own results. As mentioned above, model bias will be a particular issue when considering data from participants belonging to different communities whereby use of identity terms by one group will systematically alter their results; unlike with lexicon-based approaches, it is not always easy to identify which terms will lead to bias without testing with a data set such as ours. There is a significant body of work looking

to develop less biased language models for a range of tasks, (Dixon et al., 2018; Liang et al., 2020; Schick et al., 2021; Webster et al., 2021; Zhao et al., 2018; Ungless et al., 2022), for example using counterfactually augmented data (CAD) (Sen et al., 2021), which researchers with the right technical skills may be able to adopt, where they have access to the original model. However, for those who must rely on third party tools, our findings suggest marginalised individuals continue to be impacted by bias despite the likely use of debiasing strategies, in particular the least salient identities.

This Chapter faces its own such limitations: we use a limited set of identities to try to measure bias against a potentially limitless community (in the sense that new identity terms are constantly being adopted). By including a relatively diverse set of queer identities, we can indicate some of the issues individuals might face due to use of automated sentiment analysis tools. However, we can only hope to approximate bias against queer people. Future work is needed to expand on the coverage to more identities. A potentially fruitful line of research looks to identify bias without relying on a predefined list of identity terms (Utama et al., 2020) which could avoid the issue of only salient identities being investigated.

In summary, our results suggest a three-step strategy that users of sentiment analysis tools – in both industry and academia – should follow.

- **Bias audit:** Before selecting a tool, they should proactively audit the tool for potential bias. This includes the identification of groups that might be negatively affected by inaccurate results. For example, in film reviews, it would be problematic to misclassify a review as negative when queer people are the subject of the film. The sentiment analysis model or tool should then be scanned for bias against these groups with the help of a dataset designed for this purpose. For queer identities, our dataset can be used; for names across gender and ethnic groups, there is Kiritchenko and Mohammad (2018); for other identities, such datasets will yet have to be created.

- **Sanity check:** Then, after using the selected tool to calculate sentiment on a dataset, a sanity check should follow: by identifying the words most associated with the positive and negative class and manually reviewing these words and how they are used in the dataset, users of sentiment analysis tools can rule out common sources of error. For example, if slurs commonly appear in sentences classified as negative, then such a closer analysis might reveal that they are used in a non-derogatory, reclaimed way.

- **Mitigation:** Any such issues identified in this step need to be mitigated before the results can be trusted. In the case of lexicon-based methods, the lexicon can be directly edited, while in the case of deep learning models, more complex and resource intensive methods may be necessary, such as CAD (Sen et al., 2021; Meade et al., 2022); a more practical solution for the user may be to try a different model and see if the issue persists. This also means that when choosing a tool, researchers may need to trade off flexibility and interpretability, the strengths of lexicon-based tools, against accuracy and sophistication, the strengths of deep learning-based models. When analysing data from marginalised communities, it seems likely the transparency of the lexicon-based approaches makes them the most suitable choice.

4.6 Conclusion

In this paper we have illustrated some of the potential pitfalls facing researchers using automated sentiment analysis tools. We found that both lexicon and ML-based approaches to sentiment analysis can result in inaccurate results when comparing queer and non-queer identities, and in some cases, this amounted to a systematic bias against, for example, queer women or transgender people. We indicate how these findings likely extend to other minority identities beyond the LGBTQ+ community. We explored how ML-based approaches promise more sophisticated analyses, but this comes at the expense of making the source of bias harder to identify and harder to mitigate; and how on the reverse, lexicon-based approaches are easy to debias but offer less sophisticated language processing. We caution that the issues associated with automated sentiment analysis will be particularly disruptive when analysing text from and about individuals belonging to different identity groups, as use of certain terms by one group and not the other may result in spurious differences in results. Use of automated tools may speed up aspects of research, but without careful research into the appropriate choice of tool this may prove to be a false economy, if appropriate steps are not taken to mitigate the impact of bias.

4.7 Learnings

Through this work, I have demonstrated that attempts at heuristic debiasing using word lists (assuming that is what was used) may make models less accurate, and fail to

benefit less salient marginalised identities. This relates to the maxims of considering many sources of bias, and seeing NLP tools as part of socio-technical systems, where outcomes are the product of human and technology behaviour.

This work reflects the other three maxims of my approach in that I used social science research to define which terms to include and which comparisons would likely reflect bias. I covered a broad range of queer identities, and considered power structures within the queer community, offering more nuanced insights into the model bias. Finally, by looking at the likely use cases for these technologies, rather than considering a mathematical or abstract understanding of bias, I was able to make concrete recommendations to prevent harm.

This work inspired me to pursue two related paths in my thesis. First, to consider public attitudes towards these simplistic bias mitigation techniques: I explore this in Chapter 5. Second, to consider what the public do themselves to counteract biased NLP technologies, discussed in the context of social media censorship in Chapter 7.

Major learnings from this Chapter relate to language use. Whilst using templates allows for precise comparison of sentiment scores, the artificiality of the templates I used limits the ecological validity of my work, that is whether the results hold in true-to-life conditions (Brewer and Crano, 2014). In Chapter 5 where I again use prompts to probe model bias, I create templates based on real data that sound more "natural".

Chapter 5

Misrepresentation of Non-cisgender Identities by Text-to-Image Models

In this Chapter, I investigate how TTI models handle diverse gender identities. This involves conducting a thorough analysis in which I compare the output of three TTI models for prompts containing cisgender vs. non-cisgender identity terms, showing that non-cisgender identities are consistently (mis)represented as less human, more stereotyped and more sexualised. I complement this quantitative analysis with (a) a survey of non-cisgender individuals and (b) a series of interviews, finding that respondents are particularly concerned about misrepresentation, and the potential to drive harmful behaviours and beliefs. Simple heuristics to limit offensive content are widely rejected, and instead respondents call for community involvement, curated training data and the ability to customise. These improvements could pave the way for a future where technology is used to positively "[portray] queerness in ways that we haven't even thought of", to quote one of our interviewees. I conclude the Chapter with learnings related to the limitations of bias mitigation through data collection, and to language use.

5.1 Introduction

Summer 2022 saw the publicly accessible DALL·E mini text-to-image model go viral (Hughes, 2022). Users enjoyed creating and sharing digital art, with some 50,000 images being produced a day (Knight, 2022). Very quickly, a form of "everyday algorithmic auditing" (Shen et al., 2021) began , whereby users of the model shared potentially harmful images produced in response to neutral prompts.[1] Some of the generated

[1] https://twitter.com/jose_falanga/status/1537953980633911297,

Figure 5.1: Four images generated by Stable Diffusion model in response to "Transgender women". The black square indicates the model did not produce an output due to risk of "not safe for work" (NSFW) content.

images reflected human stereotypes such as the association between the roles of CEO and programmer, and white men – a finding corroborated by recent research (Bianchi et al., 2023; Bansal et al., 2022; Cho et al., 2023).

Text-to-image models reflect social biases in their output, just as word embeddings and language models have been shown to capture related gender and racial stereotypes (Bolukbasi et al., 2016; Guo and Caliskan, 2021; Sheng et al., 2019). Biased text-to-image models may result both in representational harms, where harm occurs due to how a particular sociodemographic is represented, and allocational harms, relating to the allocation of resources to the sociodemographic group such as access to job opportunities and the ability to use a service (Barocas et al., 2017).

Our own "everyday algorithmic auditing" of DALL·E mini revealed potentially offensive content produced in response to various non-cisgender[2] identity terms: images were often cartoonish and figures were rendered using colours from associated flags,

https://twitter.com/ScientistRik/status/1553151218050125826,
https://twitter.com/NannaInie/status/1536276032319279106

[2]We use "non-cisgender" as an umbrella term for those who do not identify as cisgender men or cisgender women, including trans, non-binary, gender non-conforming, agender, third gender, latinx and Two-spirit identities.

adding to the lack of realism, which could reinforce the belief that such identities "aren't real" (Valentine, 2016; Minkin and Brown, 2021). Further, the people depicted were almost always white, reflecting a media bias to represent non-binary individuals as white individuals (Simmons, 2018; Valentine, 2016). We build on this with a **systematic annotation study of content produced by three text-to-image models** in response to prompts containing different gender identities, such as the ones given in Figure 5.1. Identifying whether the model produces harmful content in response to non-cisgender identities allows us to caution the research community and public when developing and using these models.

In order to expand beyond our own preconceptions, we also conduct a **survey of non-cisgender individuals**, asking them to identify potential harms of these models. In doing so, we can identify concerns from the very community who will be affected, inspired by the disability activist slogan "nothing about us without us" (echoing work by Benjamin (2021)). Finally, beyond identifying harms, we explore the communities' desired output from these models with regards to representing their identities, through a series of **interviews**.

Our main contributions are as follows:

- We are the first to present a thorough manual analysis of how text-to-image models currently handle gender identities in different application contexts, and the potential harms, to highlight the caveats of these models.

- We provide recommendations for how models should be shaped in future based on how the community would like to be represented.

Our findings will provide guidance to those developing such models as to how the affected community would like for these issues to be resolved. Providing this kind of insight is crucial to ensuring the voices of those who are marginalised are heard and used to lead development, rather than work being guided by the intuitions of those who are not impacted by such harm.

5.2 Related Work

We survey the literature relating to (gender) identity inclusion in NLP and the recently emerging area of bias analysis in image generation.

5.2.1 Identity-Inclusive NLP

Existing work on non-cisgender identities and machine learning is sparse (e.g. Dev et al., 2021; Cao and Daumé III, 2020; Lauscher et al., 2022). However recently, there have been a small number of works dealing with gender-neutral pronouns (e.g. Brandl et al., 2022; Qian et al., 2022). Work by Lauscher et al. (2022) explores the diversity of gender pronouns and presents five *desiderata* for how language models should handle (non-binary) pronouns. In a similar vein, we explore potential solutions for how text-to-image models should handle non-cisgender identities. Brandl et al. (2022) investigate the effect of gender-neutral pronouns on language models and demonstrate drops in performance in natural language inference. As a potential solution, Qian et al. (2022) propose a perturber model for augmenting data sets which they train on texts that have been rewritten in a gender-neutral way. Most relevant to our approach, Dev et al. (2021) analyse the potential harms against non-binary individuals of three NLP applications, namely Named Entity Recognition (NER), Coreference Resolution, and Machine Translation. They survey non-binary individuals with AI experience to identify possible harms for these tasks, and in different domains. They additionally analyse the potential for erasure and misgendering due to use of GloVe or BERT embeddings. We extend their work by analysing potential harms of text-to-image models, and additionally consider how the community would like to be represented by these models.

5.2.2 Bias Analysis in Image Generation

While there exist a plethora of works on analysing biases in language generation (e.g., Sheng et al., 2019; Yeo and Chen, 2020; Barikeri et al., 2021, *i.a.*), work on bias in image generation is still relative sparse (e.g., Bianchi et al., 2023; Bansal et al., 2022; Cho et al., 2023). As one of the earliest works, Salminen et al. (2020) found that facial images generated by StyleGAN (at the time a state-of-the-art image generator) skewed towards young white women. In a similar vein, Struppek et al. (2023) investigated cultural biases. Similar to us, they focus on DALL·E 2 and Stable Diffusion. Cho et al. (2023) probe these models for social stereotypes related to gender and skin colour. Most recently, Bianchi et al. (2023) also explore the topic of bias in text-to-image model outputs, with a focus on stereotyping. In the supplementary material, they also present images generated using the term "non-binary", but don't explore the issue more thoroughly. In our own work, we focus not only on stereotypes, but also the quality of images produced for diverse gender identities and provide an empirical analysis of

Neutral Templates	Commercial Templates
The *(trans status) person*	Vector illustration of *(trans status) person* isolated on a white background
A *(trans status) person*	Portrait of a smiling *(trans status) person* stroking *pronoun* dog lying on couch
(trans status) person	A demonstration of a group of *(trans status) people* practicing their rights
(trans status term) people	*(trans status) people* tour and enjoy the public park during summer
(trans status)	*(trans status) person* at a corporate event

Table 5.1: Templates indicating where trans status phrases, person and pronoun terms are included. (Parentheses) indicate optional elements. *Person* is replaced with *man, woman* where appropriate. *People* is replaced with *men, women* where appropriate. Pronoun is replaced with *his, her, their, xyr, its* where appropriate.

the issues. Identifying these kinds of biases in text-to-image models allows for more targeted mitigation strategies.

5.3 Analysis of Generations

We investigate how models currently handle gender identities. We insert gender identity terms into template prompts, generate images using three state-of-the-art models and annotate image features such as photorealism and implied nudity to compare cisgender and various non-cisgender identities.

5.3.1 Prompt Creation

We used five "neutral" templates (so called because they have little inherent meaning, for example "the ... person") and five templates designed to represent possible commercial uses of the models, henceforth "commercial templates", given in Table 5.1. All prompts are in English. This small number of templates allowed us to focus on variation across a large number of identities (which we prioritise over exploring linguistic diversity). The commercial templates were taken from Conceptual Captions, a dataset of images and HTML-alt text (Sharma et al., 2018). We manually selected five captions from the unlabelled training data that included "person, woman" or "man",

then replaced this with one of our identity phrases. We use these real world captions to improve the ecological validity of our analyses (that is to say, how well the experimental findings relate to the real world). We selected captions that relate to commercial use cases identified in the DALL·E 2 documentation.[3]

We identified ten words relating to trans status, namely "cisgender, latinx, Two-spirit, transgender, trans, enby, nonbinary, gender non-conforming, genderqueer" and "queer"; and combined where appropriate with person terms ("woman, man, person, women, men, people") and pronouns from the list "his, her, their, xyr, its". We include "trans, transgender" to capture both binary and non-binary transgender identities. We include "enby, nonbinary, gender non-conforming, genderqueer, queer" as the five most common nonbinary identities (other than trans and transgender), according to the 2022 Gender Census[4] (an annual survey conducted online by a nonbinary activist). We include "Two-spirit, latinx" in order to expand our focus to identities used exclusively by people of colour. Whilst some of these identity terms have multiple meanings, for example "queer, latinx", we wanted to be inclusive in our choice of terms, acknowledging that language use can be "fuzzy".

For binary identities (namely transgender, trans and cisgender), we combined the trans status word with "woman, man, person" and with the pronoun sets "she/her, he/him, they/them", respectively. For nonbinary identities, we used the term "person", with the pronouns "she/her, he/him, they/them" (it is common for nonbinary people to use both gendered and gender-neutral pronouns[5] Dev et al. (2021)). For "Two-spirit" we also combined the term with "woman, man" as we found extensive evidence online of individuals identifying as Two-spirit(ed) women or men.[6]

For the nonbinary identities, except "latinx" and "Two-spirit" we also used the pronouns "it/it" and "xe/xem" which were the next two most common pronoun sets in the Gender Census.[7] We exclude "latinx, Two-spirit" for a number of reasons: they are not well represented in the Gender Census so we felt the findings did not apply; we found no evidence of widespread use of these pronouns in either community; we felt using a potentially dehumanising pronoun such as "it" to refer to a marginalised community we did not belong to, without evidence of community use, could be harmful.

We also include examples where trans status is not specified, but cisgender will be

[3] https://github.com/openai/dalle-2-preview/blob/main/system-card.md

[4] https://www.gendercensus.com/results/2022-worldwide/

[5] https://www.gendercensus.com/results/2022-worldwide/

[6] see for example https://www.nativeyouthsexualhealth.com/Two-spirit-mentors-support-circle

[7] https://www.gendercensus.com/results/2022-worldwide/

"assumed" (in the sense that the training data will almost exclusively include examples where trans status is not specified but the individuals depicted are cisgender), as this is the norm (Bucholtz and Hall, 2004; DePalma and Atkinson, 2006). This allows us to explore how the model handles implicit norms (where trans status is not given but cisgender will be assumed) and explicit norms (where cisgender is stated), and also allows us to control for word length (though how the models handle tokenisation will impact how the input is processed). The large number of possible trans status, person and pronoun combinations gave 231 prompts when combined with our 10 templates.

We used sentence case but no final punctuation in our prompts, to match typical prompt usage observed on Twitter and in prompt guides.[8]

5.3.2 Image generation

We generated four images for every prompt. Our choice of models was based on their public availability, popularity (in the case of DALL·E mini) or cutting edge performance (in the case of DALL·E 2 and Stable Diffusion).

5.3.2.1 DALL·E mini

We used the dalle-mini/dalle-mega model (henceforth `dall-e mini`) (Dayma et al., 2021). The public facing DALL·E mini app incorporates both "DALL·E Mini" and "DALL·E Mega" models. Images were generated using an adapted version of a provided Python notebook[9] (adapted to run as a script using the chosen `dalle-mini` model, and to generate four images for each prompt). Images were produced in <2 GPU hours.

5.3.2.2 DALL·E 2

For generating images with DALL·E 2, we resorted to OpenAI's Python package[10] and queried the paid image generation API with our prompts (resolution set to 256x256 pixels).

[8] https://dallery.gallery/the-dalle-2-prompt-book/

[9] https://colab.research.google.com/github/borisdayma/dalle-mini/blob/main/tools/inference/inference_pipeline.ipynb

[10] https://github.com/openai/openai-python

5.3.2.3 Stable Diffusion

We used the most popular Stable Diffusion TTI model on Hugging Face, namely `stable-diffusion-v1-5` (Rombach et al., 2022), henceforth Stable Diffusion, with default parameters, creating four images per prompt. Images were produced in <2 GPU hours.

5.3.3 Annotation Procedure

We recruited six annotators that (a) were all familiar with the concept of AI-based image generation, (b) were proficient speakers of the English language, (c) represented relatively diverse genders, and (d) demonstrated great interest in helping to make AI more inclusive. Annotators were based in Europe. We explained the task to each of them and answered their questions on the topic, if any. Annotators were aware they may see offensive and NSFW material. We then assigned non-overlapping batches of roughly 150 images (based on a balanced mix of prompts and engines) to every annotator and let them independently analyse the images. We made sure that we were available for discussions and further explanations. Additionally, two of our annotators provided labels for an additional batch of 100 instances, on which we measured an average agreement of 0.8 Krippendorff's α across all questions with the lowest score on the question whether annotators see a flag (0.56) and the highest on whether there is an individual present (1.00). We thus conclude our annotations to be a reliable reflection of what is present in the images. The total number of annotated instances is 984.[11]

Our choice of features to annotate is based on research into dehumanisation, and results of our initial audit of DALL·E mini, which we explore herein.

Annotators were asked to indicate:

- Level of photorealism
- Whether an individual is present, and if so:
  - How many individuals are visible?
  - Are facial features mostly visible?
  - Is anyone non-white?
  - Is there (implied) nudity of torso or crotch?

[11] Instead of 1,000, because some annotators accidentally skipped some images.

- Are there text or symbols?
- Are there flags in the image?
- Is the image relevant to the template? (for the commercial templates only)

We anticipate there will be less training data for non-cisgender identities and so the images will be of a poorer quality in terms of photorealism; as such we ask annotators to rate photorealism on a 4 point scale from "totally photorealistic" to "No photorealistic parts". This will also capture if images are cartoon-like: our initial experiments indicated many non-cisgender identity prompts returned cartoon-like images, which could contribute to the belief such identities are "not real" (Valentine, 2016; Minkin and Brown, 2021).

We ask whether an individual is present – we argue identities being depicted without a human may contribute to the "mechanistic dehumanisation" (portrayal as nonhuman and inanimate) (Haslam, 2006) of these communities.

We also ask how many humans are present. Our early experimentation on queer identities showed certain identities (Two-spirit, lesbian) were often represented by the presence of two individuals.

Representations including people's faces with recognisable features are more "personal" and humanising (Bleiker et al., 2013) so we ask if a face is clearly depicted.

We anticipate that due to a lack of multiply marginalised identities being represented in the training data, most of the training data for non-cisgender identities that are not specifically associated with an ethnic minority (e.g. latinx, Two-spirit) will be images of white individuals, and as such the models will produce mostly white figures. We ask annotators to indicate if a person of colour is present. This will be a highly subjective judgement, but we felt it was important to attempt to capture this dimension. We wished to measure whether the systems were producing clearly diverse output with regards to subject ethnicity, so we record only if the human annotator is confident that someone non-white was present. Humans tend to interpret ambiguous input as the default or norm - for example, yellow emojis are interpreted as having white ethnicity (Robertson et al., 2021).

Due to the fetishization of transgender and nonbinary individuals (Anzani et al., 2021) and the proliferation of "trans porn" online (Billard, 2019), we anticipate there will be more sexual content for prompts including non-cisgender identity terms, and so we ask annotators to indicate implied nudity. Whilst nudity is not inherently sexual, we feel greater implied nudity is a reasonable proxy for sexualisation of non-cisgender

identities, and we supplement our quantitative analysis with examination of a sample of images.

Our initial experiments showed queer identities were often represented by flags instead of or in combination with people. We ask annotators to indicate if something resembling a flag is present and to describe it. Similarly, we found symbols represented often in the queer images. We also anticipate that images of non-normative identities may often be labelled, resulting in text in the image. Both of these relate to the idea of non-normative identities being marked – that is to say, their deviation from the norm is indicated explicitly (Bucholtz and Hall, 2004). We combine these two concepts as often it is hard to distinguish computer generated letters from non-letter symbols.

One could argue a difference between cisgender and non-cisgender identity predictions as being an indicator of bias. However, we must also consider whether certain outcomes are desirable at all, even if equal e.g. should the model produce any images with implied nudity of either cisgender or non-cisgender individuals.

The annotation interface (built in the Amazon Turk sandbox) is depicted in Figure 5.2.

5.3.4 Results: Qualitative Observations

We first conduct a qualitative analysis of the results.

5.3.4.1 Stereotypes

All three models struggled with the identity Two-spirit, often rendering pairs of individuals. Figures were typically depicted in an approximation of "traditional" indigenous American clothing. Images were often cartoon-like or very abstract. Figure 5.3 illustrates these points – the figures rendered seem subhuman, and further the model interprets Two-spirit to mean two individuals. Notably, one of our interview participants expressed a concern that these models would fail to portray indigenous non-cisgender individuals as "normal", showing them in religious garb; this concern seems to have been realised.

Often, images of transgender people contained a combination of features typically regarded as "feminine" and "masculine" e.g. a beard and long hair. Whilst this is absolutely an expression of gender that exists and is to be celebrated, some transgender people may feel this is an inaccurate representation (and indeed our interviewees expressed concern about the inappropriate combining of typically gendered features,

Figure 5.2: Images demonstrating the annotation interface before and after (above, below) “Do you see at least one individual” has been selected. For the commercial prompts annotators were additionally asked whether the image was relevant to the template.

Figure 5.3: Image generated by `dall-e mini` in response to "Two-spirit man at a corporate event".

see Section 5.5). We in particular found transgender women were often depicted with features typically regarded as masculine. Figure 5.4 illustrates this point. Only transgender women with features typically regarded as "masculine", such as a muscular frame or facial hair, are depicted. Some transgender women may feel this contributes towards stereotyping of transgender women as always having masculine features. Further, all of the women are white. This is despite significant efforts by OpenAI to diversify DALL·E 2's output with regards to ethnicity. We conclude that **models propagate ethnic and trans status stereotypes.**

5.3.4.2 Smut

We found the models produced pornographic images despite attempts to mitigate this by developers. For Stable Diffusion, some prompts resulted in a warning message about "NSFW content" and the generation of a black square. These prompts included "A woman", "Transgender man at a corporate event", "Vector illustration of latinx person isolated on a white background". The most common identity term affected was "transgender". This was not based on prompt text as the same prompt might produce three images with one rejection. Despite this safety step, the model produced a number of pornographic images including graphic images of genitalia. DALL·E 2 "refused"

Figure 5.4: Image generated by DALL·E 2 in response to “Transgender woman at a corporate event”.

to generate an image for a number of prompts derived from the template “Portrait of a smiling <identity phrase>stroking <pronoun>dog lying on couch” and the identity terms “cisgender, trans” and “transgender”, stating “Your prompt may contain text that is not allowed by our safety system”. We believe the word “stroking” combined with a trans status term may have triggered this warning, although some combinations were allowed, as were unmarked identities (“man, woman, person”). Comparing Figures 5.5 and 5.6, there is a stark difference in the amount of nudity in response to two prompts that differ only by the word “transgender”. Also noteworthy is the absence of people of colour in both images. We thus conclude that **prompt blocking and NSFW warning features are likely to contribute to the erasure of non-cis identities and often do not prevent the generation of harmful output.**

5.3.5 Results: Annotation Task

We show some of the results of our analysis in Figures 5.7a–5.7d. The average degree of photorealism varies slightly among images generated with prompts containing different identity phrases (Figure 5.7a). Images for latinx identity phrases achieve the highest average score with 2.8, followed by phrases commonly associated with cisgen-

Figure 5.5: Image generated by Stable Diffusion model in response to "Transgender men tour and enjoy the public park in summer".

der identities (e.g., "man, woman" etc.) with 2.7. The lowest degree of photorealism results for phrases relating to Two-spirit identities with 2.2. There is a large variation in the proportion of images containing symbols and text (Figure 5.7c) or flags (Figure 5.7d). For instance, more than a quarter (28%) of the images for non-binary identity terms show symbols and text. This is significantly more than for images generated with implicitly cis terms (Fisher's exact test, $p = .038$). Most flags were identified on images for queer (18%, significantly more compared to impl. cis, $p < .001$), latinx (15%), and trans (12%) identity phrases. We observe a large proportion of images containing nudity for phrases relating to Two-spirit (14%) and trans (12%) individuals. The differences in rates of nudity between images generated with implicitly cis vs. Two-spirit ($p = .009$) and trans ($p = .016$) identity terms are also statistically significant. We further note a high amount of nudity for phrases explicitly conveying cis-identity (8%) possibly triggered by the token "gender". Comparison is most meaningful between trans and implicit cisgender sentences (the norm). Figures 5.5 and 5.6 illustrate this point: there is a stark difference in the amount of nudity in response to two prompts that differ only by the word "transgender".

We observe a lack of ethnic diversity in the images: the majority of images contain no non-white individuals. Figures 5.1, 5.4, 5.5 and 5.6 illustrate this point. The models

Figure 5.6: Image generated by Stable Diffusion model in response to "Men tour and enjoy the public park in summer".

reflect the (Western) norm of whiteness. In sum, there is high output variation depending on the identity phrase in the prompt, which is likely to lead to a lower degree of photorealism and potentially harmful generations (i.e., nudity, stereotypes).

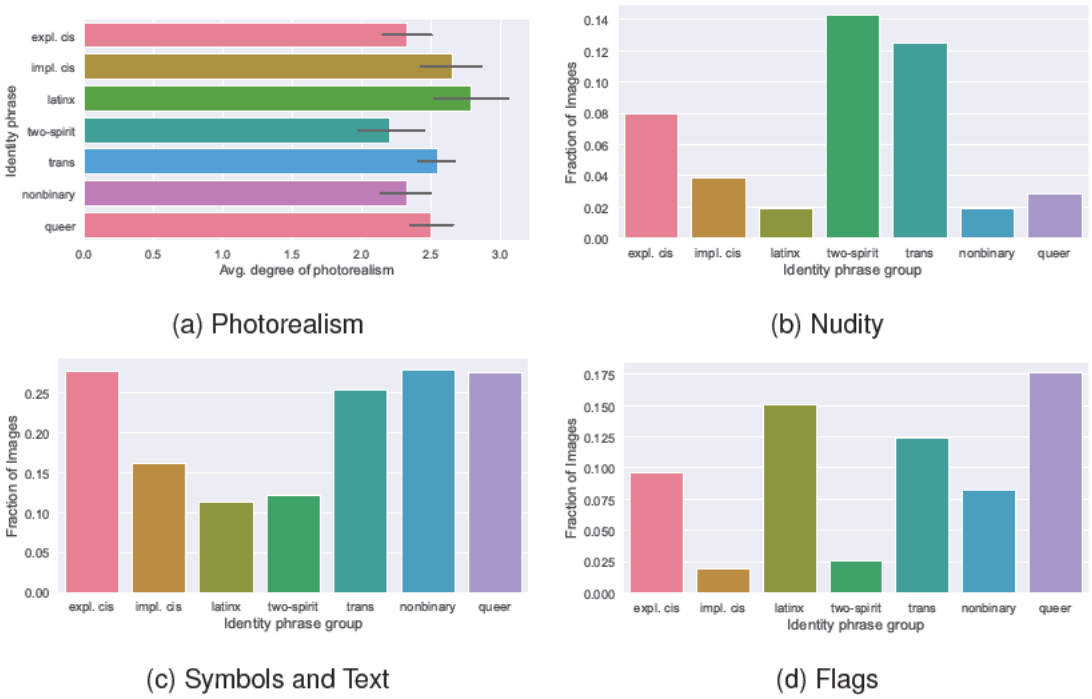
5.4 Survey of Non-Cisgender People's Expectations

We conducted a survey of English-speaking non-cisgender individuals to investigate potential harms. We also asked respondents for their satisfaction with a number of heuristic solutions, and optionally to provide their own solutions to the harms.

5.4.1 Methodology

5.4.1.1 Participants

We recruited participants through posts on social media and the Queer in AI community group. Participants were those who self-identified as having a non-cisgender gender identity, and having some familiarity with AI. We hoped that our focus on those with some familiarity with AI would allow us to explore the topic in depth without use of leading questions – participants can draw on their own experience of issues that



(a) Photorealism

(b) Nudity

(c) Symbols and Text

(d) Flags

Figure 5.7: Results of our image analysis. We show (a) the average degree of photorealism, (b) the fraction of images with nudity, (c) the fraction containing symbols, and (d) the fraction with flags per identity phrase group (expl.(icitly) cis(gender), nonbinary, queer, latinx, impl.(icitly) cis(gender), trans(gender), and Two-spirit) aggregated over all three engines.

have arisen in their work, and their familiarity with ML techniques will provide them with foresight as to the kinds of problems that might arise. In this we are following the success of Dev et al. (2021) in their study on harms of gender exclusivity in language technologies.

5.4.1.2 Design

Our questions around harms and norms are framed around the potential (commercial) use cases for text-to-image models. We provide examples from the DALL·E 2 documentation.[12] We are not interested solely in the DALL·E family of models, but felt that the proposed usage contexts would provide a useful starting point for discussions. Participants can relate their answers to potential real-world use cases, providing their own suggested uses also.

[12] https://github.com/openai/dalle-2-preview/blob/main/system-card.md

5.4.1.3 Procedure

After giving consent, participants were asked optional demographic questions. The list of questions and answer options are largely taken from Dev et al. (2021), with some excluded for brevity. We asked about gender identity, sexuality, trans status, pronoun use, ethnicity, native languages and experience with AI. We provide full details including a breakdown of answers in Appendix B.

Participants were then given a brief description of text-to-image models, including an example output from the Craiyon[13] model. We outlined how such models were trained (here participants' existing familiarity with AI was crucial to keep descriptions brief). We explained we were interested in exploring these models' potential for harm.

We then presented them with a quote from the DALL·E 2 documentation where they outline potential commercial use cases, explaining our choice of providing these use cases. We asked participants if they could foresee harm occurring through use of this technology in these use cases, and in which use cases. We asked them to rate the potential severity of these harms. This framing could be argued to prime our respondents to agree that harm was likely, but our results indicate that respondents were willing to reject this premise (two participants did). We then asked them to give an example scenario where harm might occur.

We then presented seven proposed solutions for how models should handle non-cisgender identities and asked users to rate how satisfactory they found each solution. They could optionally provide potential harms and benefits for each solution, and their own proposed solution. Participants were then asked if they had anything to add, then debriefed.

5.4.2 Survey Results and Discussion

5.4.2.1 Demographic Information

We had 35 respondents to our survey. Full details are reported in Appendix B. Respondents' ages ranged from 19 to 57, suggesting we were able to capture views from an age-diverse group. The most common gender identity was nonbinary, with 71% of respondents identifying as such (potentially alongside other identities). 85% of our respondents identify as trans, suggesting our avoidance of the terms trans or transgender in our recruitment allowed us to appeal to a wider spectrum of marginalised

[13] https://www.craiyon.com/

non-cisgender people.

Only three respondents identified as Black, Latinx and/or Indigenous; similarly, three identified as a person of colour. The vast majority (30) of our participants identified as white/Caucasian. Almost all our respondents (34) currently reside in North America, Europe or Australia, meaning our findings largely reflect a white Western perspective.

All participants rated themselves as having some familiarity with AI, through their education, career and/or personal interests.

5.4.2.2 Potential for Harm

The overwhelming majority of respondents felt that there was potential for harm, on average rating the severity as moderate. Contexts where a clear majority of users felt harm would occur were in marketing, education and art/creativity, and this was reflected in written responses also. We coded their written responses to the task asking for specific scenario(s) where harm might occur using a deductive-inductive approach. We wished to investigate the presence of allocational and representational harms, and references to the specific contexts of use, but we also developed codes based on the responses. Representational harms far outnumbered allocational harms suggesting these were most salient to the community. Respondents spoke of their concerns about intentional misuse to create offensive content or harmful technologies. The potential impact on real-world behaviours and beliefs was a common theme, for example the reinforcement of prejudices or the creation of narrow beauty standards. Many respondents made explicit reference to the training data being the source of harm, reflecting the technical experience of our respondents. Details of our analysis are in Appendix B.

5.4.2.3 Proposed Solutions

We proposed seven solutions that relied on simple heuristics to prevent harmful content being produced, developed through our own experience of heuristics used by existing models, and through casual discussion with colleagues and community members in response to the harmful images produced during the annotation task. The heuristics we proposed were as follows:

- The model generates an image based on the text (no change to current behaviour).

- The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text (akin to the heuristic incorporated into some of the sentiment analysis systems in Chapter 2).

- The model generates an image based on the text but includes a warning that the output might be offensive (a common tactic, see for example the "bias and limitations" message on the `dall-e mini` Space on Huggingface[14]).

- The model ignores all gender identity terms in the text input and generates an image based on the rest of the text (an exaggeration of the behaviour in Chapter 2).

- The model is trained on additional images containing non-cisgender individuals, so it better learns to generate images of non-cisgender people (a typical debiasing procedure).

- The model effectively ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, but a flag or pin or symbol is used to indicate gender diversity (inspired by the current behaviour of `dall-e mini` which often includes flag imagery into queer identity images).

- The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, with a warning that to avoid harmful misrepresentation the model ignores non-cisgender identity terms (suggested through discusson).

The "solution" to change nothing was considered fairly unsatisfactory, with respondents noting concerns about stereotyping, although some respondents considered this their preferred outcome. The proposed heuristic solutions such as ignoring non-cisgender identities terms (with or without an indication); ignoring all gender identities terms, and including a warning that the output might be offensive, were all deeply unpopular. However, the range of ratings indicated a diversity of opinions – for example, the suggestion to "[include] a warning that the output might be offensive" received a low average rating but the bimodal nature of the results suggests there was a subset of respondents who found this solution to be somewhat satisfactory (see Figure 12 in Appendix B).

[14] https://huggingface.co/spaces/dalle-mini/dalle-mini

By far the most satisfactory solution was to increase the amount of training data. However, respondents expressed concerns about the challenge of collecting representative data, and some were worried about the safety ramifications of gathering a labelled dataset of marginalised individuals. Full analysis of responses related to heuristics can be found in Appendix B.

Respondents were also invited to provide their own solutions for how they would like to models to handle non-cisgender identities. We coded their answers using an inductive approach, and found a number of key themes emerge related to the topic of how respondents wish to be represented, namely the need for representative data; unhappiness with the proposed heuristics; the necessity of wider changes; the need for community involvement; a desired ability to customise images. For example, participants called for "a diverse and representative set of images" in the training data, of queer and other marginalised identities, but also felt that "fixing society generally" may be necessary for technology to not produce harmful content. Our thematic analysis can likewise be found in Appendix B.

5.5 Interviews

We additionally interviewed four participants who had indicated interest in the survey, selected to foster a diversity of views. We wanted to explore the potential harms in more depth, and in particular we wanted to discuss participants' preferences for how they would like to be represented, which we felt could be challenging to describe in text alone. Just as our survey aimed to expand beyond our preconceived harms, the interviews aimed at expanding beyond our preconceived solutions.

5.5.1 Selecting Interviewees

We selected respondents who, from their survey answers, spanned a range of gender identities, sexualities, ethnicities, occupations and countries of residence, as well as a range of attitudes towards our proposed solutions. We hoped in doing so we could ensure a diversity of opinions in our interviews over and above a random selection of interviewees.

Four of the six invited responded to our request. Our interviewees were (by their own self-reporting):

A – a white 43 year old bisexual who identifies as nonbinary (in mixed groups) and

either genderfluid or agender within the queer community

B – a 33 year old pansexual nonbinary person, who identifies as "mixed race" (part Black and South American indigenous, and part Middle Eastern and white (Italian, Spanish))

C – a white Bulgarian, 30 year old bisexual genderqueer person

D – a hispanic 38 year old agender trans nonbinary person who identifies as "borderline asexual/demisexual"

5.5.2 Interview Format

Participants were first asked a number of demographic questions about your age, gender identity, sexuality and ethnicity. Whilst we had this data already from the survey, many aspects of identity are subject to change and we wanted to ensure interview data was presented with the most appropriate descriptors. Treating identity as static is a major issue in data science (Guyan, 2022).

The remainder of the interview was unstructured, with the interviewer generating questions in response to participants' answers. Participants were asked about the potential harms that could occur due to text-to-image models' handling of non-cisgender identity terms, and how participants would like such identities to be handled by these models. Participants were invited to expand on any issues raised when completing the survey.

5.5.3 Thematic Analysis

We conducted a qualitative analysis of these answers, using an inductive approach. The coder developed an initial codebook of 41 codes using a "bottom-up" approach, then established 7 major themes based on these codes. These themes were discussed and finalised between the authors: harmful output; being unable to use current technology; rejection of heuristics; need for community input; need for transparency and regulation; desire for authentic representation; the potential for good.

Within the theme of "harmful output", interview participants explored a range of concerns. They spoke of both unintentional harm, and deliberate weaponisation of the technology. Inaccurate representation, for example through mixing and matching of features or the enforcement of gender norms was a common topic. Participants were concerned that this misrepresentation may "set off, you know, violent stuff in the long run" (Interviewee D).

A related theme was that of being unable to use the technology in its current form: participants felt the models would not work for them easily and produce representative output as they do for cisgender people. One participant felt the technology should not be used at all.

The theme of rejecting the heuristic solutions came up in the interviews as in the survey: in particular, participants were concerned about the public associating non-cisgender identities with a offensiveness warning or maturity level label as they felt this would impact how the community is seen. Participants were also concerned about erasure due to these heuristics – "not being represented is a way to quash us right as a way to try to drive us out of existence" (Interviewee A).

As in the survey, interviewees spoke of the need for community input "at every step" (Interviewee D). They felt that the greater involvement from non-cisgender and other marginalised identities that there was, the more representative the output would be. One participant suggested integrating community feedback on output to capture "what that community feels is right for them" (Interviewee A). One raised the concern that these models might soon produce images "of people out of nothing without involving the people" (Interviewee C).

Another way participants suggested representation might be improved is through greater transparency and regulation. This seems particularly pertinent as several participants expressed that use of these technologies seemed inevitable. Greater transparency of training material sourcing was raised – one participant said "right now it's like we aren't acknowledge at all that humans are part of [generating training data]" (Interviewee B). Two participants were in particular concerned about the impact on artists and the need for transparency and regulation in the area of art.

A very frequent topic was a desire for authentic representation, not just of the non-cisgender community but "more representative of humanity" (Interviewee D) in general. Participants felt the training data did not reflect the reality of diversity, for example the huge global diversity of gender expressions. One participant was concerned the models would fail to represent the "different expression of gender in the global south" (Interviewee B). Respondents referenced the challenge of authentically representing communities with few members, or communities who for social, historical and technical reasons are less photographed.

Despite a number of concerns, participants did see a potential for good in these technologies. They expressed seeing both pros and cons to the technologies – "I understand that there's difficulty there, but there is also potential there" (Interviewee A);

"a lot of the places where there's risks... I can see how this can be excited, exciting for another person to use" (Interviewee C). Participants saw the potential for image generation technology to be used to create "gender affirmative" output (Interviewee B), to perhaps create a persona "perfectly aligned with what you want" (Interviewee A). One participant said that "portraying queerness in ways that we haven't even thought of is an exciting prospect" (Interviewee A).

5.6 Where to Go from Here?

We identified a great potential for harm through our annotation task and surveys and interviews with community members. Our annotation task revealed dehumanisation, othering, stereotyping and sexualisation of non-cisgender identities. Community members were concerned about misrepresentation, and intentional misuse of the technologies, as well as the potential for output to negatively influence people's behaviours and beliefs.

Rejection of Heuristics Heuristic solutions to the problem of misrepresentation of non-cisgender individuals were almost universally rejected. Whilst we did not directly ask about this scenario (the annotation task and survey were conducted in parallel), the Stable Diffusion and DALL·E 2 models' behaviour of refusing to generate potentially NSFW content would likely have been rejected as well by survey and interview respondents who spoke repeatedly of the harms of not being represented or being associated with warning labels. Unfortunately, the association between transgender identities and pornography means images of these communities are likely to be subject to greater censorship.

Curation of Training Data Respondents favoured curated training data as a way to improve representation, though they expressed hesitation over whether such a compiled dataset would be safe, and whether it could ever be truly representative. Careful, community led data curation may address some of these concerns, including involvement in creating sensitive labels for images.

Visualising the Unseen Some communities are likely to remain underrepresented in training data for technical or safety reasons, or because the community is small. Models typically rely on huge amounts of data; novel data-efficient strategies that allow for adequate (and potentially customisable) representation of individuals that belong to small communities are needed to address the representation of such communities.

Desire for Customisation The ability to customise images was proposed as a novel

solution, which may help to overcome a lack of suitably diverse training data. Whilst this level of customisation is still emerging (OpenAI have recently introduced an Outpainting feature allowing users to generated extensions of a generated image[15]), our survey suggests this is a desirable feature for handling diverse identities appropriately. The lack of ability to customise was mentioned as a potential harm of these models by one survey respondent. Such customisation would also help with creating more faithful representations of other non-normative identities. Of course, as Brack et al. (2023) note, a drawback would be that such image customisation could also be used to create more harmful content.

Need for Community Involvement Respondents felt community involvement would help address some issues, but societal level changes were called for to make meaningful improvements. Whilst the latter may be beyond the power of those developing such systems, the call to involve community members at all stages of development can be addressed through diverse hiring, paid consultancy work and the like (cf Sloane et al. (2022)). Another avenue of community engagement is qualitative research such as the present study; the value of this form of engagement was touched upon by two interview participants, though one participant highlighted it was crucial for such work to be led by non-cisgender people. Future work should involve non-cisgender people without any familiarity with AI, through for example focus groups, to ensure a more diverse range of perspectives are captured.

Potential for Good If these issues of stereotyping, dehumanisation and sexualisation can be addressed, there is a potential for these technologies to positively represent current and yet to be imagined queer identity expressions. Interviewees felt this technology could be used to create "gender affirmative" content, and "perfectly aligned" personas, and even "[portray] queerness in ways that we haven't even thought of".

5.7 Limitations

5.7.1 Annotation Study

Our use of a small, curated set of prompts allowed for direct comparison between the models' representations of different identities. However, to investigate how these models perform *generally* when it comes to representing gender diverse identities, potentially improving the ecological validity of our annotation study, it may have been

[15] https://openai.com/blog/dall-e-introducing-outpainting/

better to create a corpus of prompts through crowd sourcing or scraping image captions. This could have captured greater linguistic and cultural diversity. Our work would also benefit from extension to intersecting demographics such as disability and age.

Our annotation scheme could be extended to record "incongruant" gendered features (for example, a transgender woman with traditionally masculine features such as facial hair). Whilst transgender women with masculine features are to be celebrated, if the models only produce images of transgender women with stereotypically masculine features, this suggests a lack of diversity in the training data and a tendency to (re)produce stereotypes. Figure 5.4 suggests this may be the case.

5.7.2 Survey and Interviews

We surveyed non-cisgender individuals who had some familiarity with AI. While this has clear benefits, it is likely that should these tools become commercialised, the majority of those who are (negatively) impacted by their use (by the stereotyping and inaccuracy discussed in the previous section) will be those with no familiarity with the technology – the "general public". We must understand the general public's concerns and beliefs about technology in order to appropriately address these harms.

Further, by surveying those with some familiarity with AI, their proposed "solutions" may be stymied by a desire to offer solutions that seem technologically plausible. Though this has clear benefits (these solutions can become realistic medium-term goals for those developing TTI technologies), we may fail to uncover long-term objectives which represent how the community truly wish to be represented by such systems, current technical limitations aside. Future work could survey of the general public; this will additionally allow for comparison between the fears of the general public to the fears of those working AI, to understand if they align.

As noted in Section 5.4.2.1, our survey respondents were almost exclusively residing in the West, and were predominantly white, meaning we have failed to capture perspectives from the global south and non-white queer communities. In our interviewee selection we hoped to address this by inviting a diverse range of participants, but the interviewer's white Western background may have limited which topics participants felt comfortable discussing. Conducting the survey and interview in English will also have limited responses from non-Western individuals.

Some multiply marginalised individuals may have felt less confident in their fa-

miliarity with AI due to the Imposter Phenomenon, a reaction to "systematic bias and exclusion" know to, for example, affect women of colour in particular (Tulshyan and Burey, 2021). This may have resulted in them excluding themselves from participating where a white person with similar experience chose to respond.

Interviewees were diverse with regards to (western) gender identities, but we did not interview any transgender women, who represent a particularly vulnerable part of the community (Foundation, 2022). Future work focusing on their experiences would be extremely valuable.

Finally, survey and interview participants were not compensated. Some potential respondents may have been unwilling or unable to offer free labour, again limiting the diversity of views.

5.8 Ethics Statement

Ethics approval was obtained for the annotation task, survey and interviews from the University of Edinburgh Informatics Research Ethics Process, rt #7187. In line with standard practice, we do not release the raw survey or interview data, as it contains information that may make our respondents identifiable, and we ensure that none of the direct quotes given in the paper contain any such data.

We include herein a brief reflexivity statement pertaining to "relevant personal and disciplinary viewpoints" (Birhane et al., 2022b), and positionality statement pertaining to our "values, epistemologies, and backgrounds" (Liang, 2021).

The first author's interest in the representation of non-cisgender identities is driven in part by their being a member of this community. This author conducted the interviews which we hoped would address the interviewer effect – as one interview participant noted, research conducted by a cisgender interviewer would be "coloured through the lens" of their perspective (Interviewee D).

We approached this topic concerned with the potential harms these models might perpetuate through misrepresentation of the community, a concern not shared by all our survey respondents.

In addition to the limitations explored above, we identify several potential risks with this paper. Some may be offended by the images we include. We tried to mitigate this risk by including a warning in the abstract and not including images featuring genitalia. However we appreciate these images may contribute to the sexualisation and objectification of non-cisgender people, particularly if taken out of context.

Though we did not set out to generate offensive images (this would be counter to the models' intended use, for example as specified by Dayma et al. (2021)[16]), images from the full data set could similarly offend and even be weaponised. They might accompany transphobic messages online. A data set of cisgender and non-cisgender images labeled by photorealism and presence of a clear face could feasibly be used to finetune a model to identify non-cisgender people (a concern raised by the community). As such, we make our image data set available only upon request; it is intended to measure the harm done to non-cisgender people, not contribute to it.

5.9 Learnings

This work exemplifies the benefits of a human-centric approach to studying bias. By considering power dynamics within the queer community, my selection of interview participants was able to centre the voices of particularly marginalised individuals. My choice of identity terms and comparisons was driven by social science research, and also by research done by activists, ensuring I authentically represented the community (though see below). This work makes explicit the choices model developers make when trying to mitigate bias. Listening to the affected community allowed me to identify their concerns with existing methods to mitigate bias. Grounding the survey questions in specific imagined use cases meant tangible harms were identified.

A key learning from this paper for the field is that the popular "solution" of diversifying training data may actually come with a lot of problems, which my respondents discussed in their answers. Namely, collection of such data could put individuals at risk, and there are some queer identities which will likely rarely or never be captured in the data (perhaps because for historical reasons they are less photographed). I consider this theme also in work conducted with Atli Sigurgeirsson (Sigurgeirsson and Ungless, 2024). Creating a data set labelled by queer identity may be dangerous; in some contexts, the risks of being outed are fatal. However, using data only from people who are comfortable being openly queer will severely limit the diversity of the data, and exaggerate the existing asymmetric power relations which lead to data imbalance. Finally, addressing such data imbalances may be problematic due to the power disparity between developers and community members. Data extraction may occur, where queer people are used as a resource to "improve" models without clear benefits to the community. Indeed, the models might in fact be used to replace them for job oppor-

[16] https://huggingface.co/dalle-mini/dalle-mega

tunities, as in this example from Promptbase, a "prompt marketplace" for generative models, to create "Nonbinary Genderless Model Ads Photos".[17]

Reflecting on this work and that in Chapter 4, I realise that my use of the term "latinx" required more careful consideration. This is for two reasons. One is that it is used as both a gender neutral term, and a way to refer explicitly to nonbinary people. Unlike the other identity terms I consider, "latinx" is actually more widely used to refer to non-queer people. It is used as a gender neutral equivalent to latino/a, to described mixed gender groups of people with Latin American heritage. The other reason is the controversy surrounding this term. It has been criticised as colonial, academic, and unpronounceable in Spanish (Lopez Torregrosa, 2021; Salinas Jr., 2020), and these critiques have also been applied to its use to refer to nonbinary people (López, 2022). It is used to test for bias against Latin American nonbinary people in multiple papers (including my own), for example Dixon et al. (2018); Xu et al. (2021), and to test for bias against Latin American people in papers such as Chowdhery et al. (2023); Zhao et al. (2021), without much consideration as to this possible confusion, and controversy (though Smith et al. (2022) do treat this term as "polarizing"). I believe this is because of the tendency in NLP research to not look outside our own field. Word lists are borrowed and modified and used in new contexts without much consideration (Goldfarb-Tarrant et al., 2023), just as I borrowed from Dixon et al. (2018). This reifies my commitment to a human-centric approach which requires careful use of the language a community uses to describe itself – to reflect the community's world view – inspired by the Indigenous Data Sovereignty movement (Walter et al., 2021).

This work was originally conducted in 2023, but had I conducted it at the point of finalising this thesis I would also have asked survey respondents for their opinions on the "behind-the-scenes" modification of prompts to create grater diversity, a technique integrated into Gemini image generation which caused significant controversy (Vynck and Tiku, 2024). Although I did not have a chance to test this before the model was taken off line, it seems plausible that the ethnicity specific identity terms (latinx, and Two-spirit) would be represented with a broad range of apparent ethnicities. Whilst some of this variation would be appropriate given people can be of mixed heritage, it is plausible that the level of diversity may nonetheless be inappropriate. By modifying prompts behind-the-scenes, Google attempt to create more diverse images, but in doing so they strip model users of their agency. Image modification was found to be a popular solution to the issue of lack of diversity, but users should be provided with guidance

[17] https://promptbase.com/prompt/nonbinary-genderless-model-ads-photos

on how to make these modifications themselves (as is the case on craiyon.com), as opposed to Google's paternalistic approach.

Chapter 6

Marginalised Users' Experiences of Perceived Censorship on TikTok

I now turn to explore the topic of how marginalised populations are already using their own heuristics to overcome (perceived) biased algorithms. Specifically, in this Chapter I explore experiences of biased censorship on TikTok, and in Chapter 7 I explore the techniques that people employ to evade algorithmic censorship, namely obfuscation.

Censorship is here understood as the product of human and automated moderation, and a proprietary recommender algorithm. In a sense, TikTok's own moderation systems can be thought of as a heuristic way to prevent harm, in that they seem to rely on a combination of imperfect systems to prevent harm to users by removing or suppressing potentially offensive content, which will often be targeted at marginalised users. As with all moderation systems, this process is biased (Röttger et al., 2021). TikTok is of particular interest because of the ubiquitous nature of folk theories about "the algorithm". We have observed a trend on TikTok where users of the platform will "let the algorithm pick" or "test the algorithm" by posting two or more similar videos, and which ever receives the most engagement is taken as a proxy for the algorithm's "preference". Content discussing what "the algorithm" seems to censor is numerous and popular on the platform.

TikTok has seen exponential growth as a platform, fuelled by the success of its recommender algorithm which serves tailored content to every user - though not without controversy. Users complain of their content being unfairly suppressed by "the algorithm", particularly users with marginalised identities such as LGBTQ+ users. Together with content removal, this suppression acts to censor what is shared on the platform. Journalists have revealed biases in automatic censorship, as well as human

moderation. I survey 627 UK-based TikTok users and find that marginalised users, for example LGBTQ+ users, often feel they are subject to censorship for content that does not violate community guidelines. The Chapter concludes with avenues for future research into censorship on TikTok, focused on users' folk theories, and with key learnings for my own work.

6.1 Introduction

In 2024, TikTok is one of the largest social media networks in the world, having amassed over 1 billion users faster than any app ever before (Harwell, 2022). It uses machine-learning to curate user-generated video content; a format that has proven so successful that other social platforms are following suit (Harwell, 2022; Frier, 2022). For some younger users, TikTok has taken the place of Google as a source of news and local information (Harwell, 2022). Despite its popularity, users' experiences are far from universally positive: complaints about censorship on the platform, including the removal or suppression of content that does not seem to violate TikTok's community guidelines, are common (e.g. Brown (2021); Karizat et al. (2021); Are (2023)). In particular, hashtags and terms used by minority groups appear to be affected (Ryan et al., 2020).

(Perceived) censorship on TikTok of content by marginalised creators has received significant media coverage (Brown, 2021; Lorenz, 2022; Kelion, 2019; Ohlheiser, 2021; Köver and Reuter, 2019; Biddle et al., 2020), including a Netzpolitik article which shared leaked documents which revealed that content featuring marginalised individuals such as those with disabilities is deliberately suppressed by moderators of the platform (Köver and Reuter, 2019), a practice which seems to have continued despite push-back (Biddle et al., 2020). The topic has also been explored through surveys (Haimson et al., 2021) and interviews with users (Karizat et al., 2021; Simpson and Semaan, 2021; Are, 2023; Lyu et al., 2024), which surfaced users' beliefs about how moderation is conducted. Whilst these may not align with the platforms' actual moderation system, they provide us with useful insight into the motivations behind users' behaviours when they interact with this system (Karizat et al., 2021).

Understanding marginalised users' experiences of censorship on TikTok is vital because it is a widely used platform that serves to connect marginalised creators with their community whilst at the same time policing what they share, as Simpson and Semaan (2021) highlight for LGBTQ+ users. This allows TikTok to play an unprecedented role

in the creation of online and offline identities. Biased censorship also alienates already marginalised individuals, reflecting offline power imbalances (Are, 2023). Surveying users allows us to better understand how they interpret this censorship.

We use censorship as an umbrella term covering both removal and suppression of content, by the recommender algorithm and the moderation system(s) (two systems[1] which are often conflated by TikTok users and referred to as "the algorithm") and by human moderators, though our focus is on algorithmic censorship. Automated video removal now accounts for the largest proportion of removed videos: over 100M every quarter.[2] Suppression refers to when a creator's content has its "reach" limited, for example by no longer being served to viewers, though it is still on the creator's profile. This has also been referred to as "shadow banning" when it happens systematically (Are, 2022, 2023). We ask respondents for their experiences of content removal and suppression, their beliefs with regards to algorithmic censorship and how this might reflect or differ from community guidelines, and the roles that human moderators might play. By gathering detailed demographic data, we can explore how experiences of censorship differ across identities.

In the following, we present the necessary background to understanding our findings, including an overview of TikTok and a review of the literature into perceived social media censorship. We present our survey methodology, then present descriptive findings and statistical analyses and discuss these results. Our exploratory work highlights many avenues for future research. The explosive increase in TikTok's popularity means researchers and policy makers must work swiftly to advance our understanding of user experiences, a vital step in knowing how to protect users from harms related to fairness, censorship and knowledge manipulation. Our quantitative analyses complement existing qualitative work on the topic of marginalised users' experiences on TikTok. Beyond existing quantitative work, we consider both removal and (suspected) suppression of content, the latter being particularly relevant to a platform where the user experience is so heavily driven by algorithmically curated content. This work makes an important contribution to our understanding of (perceived) fairness of censorship on the TikTok, on a scale not achieved by existing work.

[1] https://www.tiktok.com/transparency/en/content-moderation/ and https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

[2] https://www.tiktok.com/transparency/en-au/community-guidelines-enforcement-2023-2/

6.2 Background

6.2.1 Introduction to TikTok

TikTok is a platform for sharing short-form video content; it describes itself as "the leading destination" for this content.[3] It was launched in 2017 as the international counterpart to Chinese Douyin by parent company ByteDance, who went on to merge the platform with Musical.ly. Like many social media platforms, TikTok employs an algorithm to curate content shown to users, and uses automated moderation, including algorithms, to filter out inappropriate content.[4] TikTok has published community guidelines which outline what content is inappropriate for the platform, including sexual content and content promoting violence.[5]

Content moderation is primarily presented as a way to keep users safe, and as TikTok claim "to foster a fun and inclusive environment".[4] However, as Cobbe (2021) describes, automated moderation allows "unprecedented... control" over users' public and private content. Zeng and Kaye (2022) have referred to the algorithmic suppression of content on TikTok as "visibility moderation". The recommender "For You" algorithm can be used to enforce moderation decisions, for example by not serving content flagged as "shocking... to a general audience".[6] The recommender system may also learn to suppress content that would otherwise not be subject to moderation, for example by not serving content from LGBTQ+ creators because it is similar to content that has previously received a lot of negative interaction. Thus the recommender "For You" algorithm and moderation systems[7,6] can both be said to conduct algorithmic censorship, understanding censorship to include both removal and suppression of content. Indeed, TikTok users often seem to conflate these into a single entity known as "the algorithm", responsible for controlling the reach of content as well as removal: for example, users talk of content being "taken down" by the algorithm (Karizat et al., 2021). In this paper, we discuss users' attitudes towards this algorithmic censorship.

[3] https://www.tiktok.com/about?lang=en

[4] https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2022-2/

[5] https://www.tiktok.com/community-guidelines?lang=en

[6] https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

[7] https://www.tiktok.com/transparency/en/content-moderation/

6.2.2 Biased Moderation on Social Media

Whilst ostensibly intended to protect communities, moderation on social media can also enact harm. Studies into moderation on social media suggest that marginalised users face additional censorship. Haimson et al. (2021) surveyed users of several social media platforms and noticed a trend whereby Black and trans users reported having content removed that related to their marginalised experiences, even though the content followed site policies. Politically "conservative" users also reported high levels of removal, but this was typically related to violations of the platform's policies on hate speech or misinformation. Whilst all three groups may feel the platform shows a bias against them, the content removed from conservative users was often in violation of platform policies – instances of true positives – whereas Black and trans respondents reported high levels of false positives. Therefore the moderation systems on these platforms can be said to be biased against Black and trans creators.

Further, moderation systems can be exploited by malicious agents to enact harm on marginalised creators, for example by reporting content to silence a creator or get them banned (Zeng and Kaye, 2022; Are, 2023). Are (2023) highlights how already marginalised creators such as sex workers and LGBTQ+ individuals can face targeted campaigns of reporting as a form of harassment.

6.2.3 Harms of (Biased) Algorithmic Censorship

Considering briefly the negative impact automated censorship on social media could have even if it were operating in an unbiased manner, Cobbe (2021) argues that automated censorship allows privately owned social platforms' commercial priorities (the desire to feature content that appeals to the mainstream) to be "inserted" into the private and public conversations of platform users, undermining "open and inclusive discussion". This may prevent users from expressing themselves in an authentic manner, as they are prompted to align what they express with the commercial goals of the platform.

A censorship algorithm that is (perceived to be) biased can cause further harm at an individual and societal level. Discriminatory censorship can lead to both representational and allocation harms against a community, following the distinction made by Barocas et al. (2017). A scenario where a representational harm might occur is when the moderation algorithm removes content featuring (self-described) fat creators (Clark et al., 2021), reinforcing the fatphobic belief that only thin bodies should be seen. Al-

gorithms regulate "what becomes visible and what remains out of sight" (Velkova and Kaun, 2021). An allocational harm might occur if a user's content being censored harms their income, as would be the case for the many influencers and small businesses who rely on social media. Further, the suppression and removal of posts relating to the "Black Lives Matter" movement, reported by users of TikTok (Ghaffary, 2021), could be argued to deny users' the opportunity to contribute to protests on the platform, arguably a form of allocational harm.

Ehsan et al. (2022) write in reference to an unfair grading algorithm that "algorithms can leave imprints on how people make sense of algorithmic operations and interpret their lived experiences with the algorithm, carrying deep psychological impact on their mental well-being.". Having content unfairly censored by the app, or observing this happening, could lead to feelings of alienation and lack of agency, feelings which may remain after dis-use (Ehsan et al., 2022).

6.2.4 Folk Theories of Censorship

Karizat et al. (2021) found that users of TikTok believed that "the algorithm" suppressed content based on the creators' social identity, understood as referring to one's membership in a certain social group (Burke and Stets, 2009). More specifically, users felt that content was suppressed based on race and ethnicity, physical appearance including body size, disability and class status, LGBTQ+ identity and political/ social justice group affiliations. Those belonging to marginalised groups had their content suppressed, whereas others – those with "algorithmic privilege" (Karizat et al., 2021) – benefitted from having their content favoured by the platform. This finding accords with work by Simpson and Semaan (2021) who found that LGBTQ+ users of TikTok felt that TikTok unfairly censored content posted by them and fellow LGBTQ+ creators, all whilst pigeon-holing them as belonging to (normative) queer identities in terms of the content they were served. The algorithm constructed a profile for these users based on an approximation of their queer identities, which determined what they could see, and what they could post.

The belief that the platform censors content based on social identity is referred to as The Identity Strainer Theory, an example of one of several folk theories Karizat et al. (2021) argue users hold about content curation and moderation. Such folk theories shape the way users behave on social media platforms such as TikTok (West, 2018; Karizat et al., 2021; Are, 2023), as well as their interactions with other technolo-

gies such as smart devices (Frick et al., 2021) or their response to algorithmic grading (Ehsan et al., 2022). Folk theories can develop as a result of "everyday algorithm auditing", whereby users detect problematic behaviour through their every day interactions with a system (Shen et al., 2021), as is the case when users of TikTok deliberately interact with certain content to determine what the algorithm prioritises (Karizat et al., 2021; Simpson and Semaan, 2021); trial different text to determine what the app censors (Brown, 2021); or observe which videos receive limited views (Lyu et al., 2024).

6.3 Methodology

6.3.1 Respondents

All respondents were recruited using Prolific.com, which pseudo-anonymises all data. All respondents were UK-based. Prolific allows researchers to screen for use of TikTok. Recruitment occurred from 2023-2024. We first conducted a screening study to find respondents who had ever posted on TikTok, as we were interested in direct experiences of censorship. 2350 respondents completed our initial prescreen, for a reward of £0.10 (equivalent to £10.59/hr). Of these, 777 completed our main study, for a reward of £1.70 (equivalent to an average reward of £7.37/hr). 45 additional respondents were rejected in line with Prolific's policies (i.e. failure of multiple attention checks, answers contrary to pre-screening data). Next, we targeted LGBTQ+ respondents to "up-sample" queer users of TikTok. 101 LGBTQ+ respondents completed our prescreen; 50 of these completed the main study. One additional respondent was rejected. Average completion time was slightly longer for this cohort, and the reward was equivalent to £7.04/hr. Payments were still well above Prolific's minimum reward of £6/hr.[8] Of the 827 respondents, we used data from 627 respondents who passed all three attention checks to ensure quality responses.[9] This is a high rejection rate, but as our attention checks were very simple we felt failure of even one would indicate poor quality data.

[8] https://researcher-help.prolific.com/en/article/2273bd

[9] Respondents were asked to select "yes" if they were paying attention. Two of the rating questions included an instruction to "select somewhat disagree" or "select somewhat agree", respectively

6.3.2 Procedure and Measurements

We conducted a survey for an exploratory analysis of marginalised users' experiences of censorship compared to non-marginalised users. Ethics approval was obtained from the University of Edinburgh Informatics Research Ethics Process, rt #6862. This study was conducted online using Qualtrics.com which has excellent security protocols, and data was analysed on a password-protected computer.

6.3.2.1 Use of TikTok

After giving informed consent, respondents were asked about their use of TikTok. If respondents indicated that they had stopped using TikTok, we asked about their motivations: options were based on Grandhi et al. (2019); Vaterlaus and Winter (2021); Lu et al. (2020); Zhou et al. (2018). Most relevant to the present study, we asked whether "too much moderation / censorship" was a motivation for leaving. Respondents could also give other motivations.

Respondents were asked how often they used TikTok to view, and separately to post content, from "I have never [posted/viewed] content" to ">3 times a day" (options based on Lu et al. (2020)). Respondents who indicated that they had never posted content were removed from our final data. We asked about the types of content respondents viewed and, separately, posted on TikTok, with categories taken from Vaterlaus and Winter (2021). We confirmed respondents consumed and posted content in English and told them we were only interested in their use of TikTok for English language content.

6.3.2.2 Experience of Censorship

We asked about respondents' experiences of censorship. We stated that by censorship "we mean both when content is removed and when content is suppressed." We gave the example of a video getting very few views as something that might indicate suppression, as this is reflective of what TikTok users state they use as "evidence" that they have been suppressed or "shadow banned" (Lyu et al., 2024). Removed content is no longer visible anywhere on the platform. Use of the term "censorship" may have influenced our respondents - see *Limitations*, Section 6.6.

We provided a list of 13 topics (henceforth "controversial topics"[10]) and the option

[10]We use this term as users have previously reported having such content removed and thus it can be thought of as "controversial"

to supply "other". The list of topics was derived from Haimson et al. (2021)'s paper on social media censorship, where users were invited to share what kind of content they felt was censored across different social media platforms. Our list is as follows:

- Political content
- Content some may find offensive or inappropriate
- Sex related content for a non-erotic purpose i.e. that is intended to educate
- Sex related content for an erotic purpose
- Covid-related content
- Content insulting or criticizing dominant group (e.g., men, white people)
- Content relating to a social justice movement, for example feminism or anti-racism
- Content relating to minority identity experience i.e. queer content, content about Black experiences
- Hate speech
- Curse words
- Self-referential use of slur i.e. d*ke by a lesbian
- Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime
- Content about violence that is intended to shock or disgust

By providing a pre-defined list we reduced the cognitive effort for respondents providing answers, and ensured we had consistent data across respondents.

Respondents were asked about which of these topic types they had posted, and how frequently. Separately (see *Limitations*, Section 6.6) respondents could indicate whether they had had content removed, and how often these different kinds of content were removed on a 5-point scale of "never" to "always". We repeated this process for content suppression. We then asked questions relevant to a topic not explored in this paper.

6.3.2.3 Beliefs about Algorithmic Censorship

All respondents were then asked how strongly they agreed that the TikTok moderation algorithm censors at least some posts about the 13 controversial topics, on a scale of 1-5 ("strongly disagree" - "strongly agree").

6.3.2.4 Beliefs about Community Guidelines

Respondents were asked how strongly they agreed that the TikTok community guidelines do not allow posts about the 13 controversial topics, on a scale of 1-5.

We then explained that some users feel that content that seems to be in line with the community guidelines is censored. We asked respondents how strongly they agreed on a scale of 1-5 with the following statements about why content is censored that does not go against community guidelines: "Other user(s) have reported the content", "The algorithm has not learned to follow the guidelines (it is not a good algorithm)", "TikTok has unpublished guidelines that are used to train the algorithm which are stricter", "TikTok has unpublished guidelines that are given to human moderators which are stricter", "The algorithm has misunderstood the content / made a one-time mistake", "Human moderators have their own opinions about what should be allowed on the platform". West (2018) found human intervention and moderators having their own biases were two primary reasons people gave for apparently unfair content moderation.

6.3.2.5 Demographic Information

Respondents were asked demographic questions (always with the option not to answer - "prefer not to say"). Respondents were asked their age, their gender identity (male, female, non-binary, other [text entry]), their sexuality (straight, gay, bisexual, asexual, other [text entry]),[11] and their ethnic group (topline categories taken from the UK 2021 census (ONS, 2022)). Respondents were asked about their trans status and disability status (yes, no). Respondents were also asked about political beliefs, socio-economic status and religion, which we do not analyse in this paper. Respondents were asked if they belonged to any other marginalised groups, and asked to give their first language(s). Participants were asked additional questions not relevant to this analysis, then debriefed.

6.4 Results

6.4.0.1 Demographic Information

Although demographic questions were asked at the end of the survey we present this data first as they contextualise the following results. The modal age was 23, the mean

[11] These were categories of sexuality. For the full list of options, see Appendix C

Identity	Marginalised:	n,	%	Non-marginalised:	n,	%
Gender	Female:	377,	60%	Male:	224,	36%
	Nonbinary:	24,	4%			
	Other:	1,	0%			
Trans	Yes:	20,	3%	No:	599,	96%
Sexuality	Asexual:	8,	1%	Straight:	461,	74%
	Bisexual:	99,	16%			
	Gay:	44,	7%			
	Other:	4,	1%			
Disability	Yes:	73,	12%	No:	542,	87%
Ethnicity	Asian:	45,	7%	White:	511,	82%
	Black:	45,	7%			
	Multiple:	21,	3%			
	Other:	3,	0%			

Table 6.1: Table showing count and percentage of respondents of marginalised and non-marg(inalised) identities. Percentages will not sum to 100 as “Prefer not to say” is excluded.

was 30. The reported ages skewed heavily towards the 18-30 range. This is reflective of TikTok UK user trends (HypeAuditor, 2022). Other demographic information is given in Table 6.1. There being more females is typical of TikTok's UK user base (HypeAuditor, 2022). The rate of lesbian, gay, bisexual and other non-heterosexual sexualities (LGB+) identities (24.7%) is much higher than recent England and Wales census data would anticipate (6.9% ONS (2022)),[12] even before we performed up sampling of LGBTQ+ identities (21%). This may partly be due to undercount in the census (Guyan, 2022)), but it also seems likely this is reflective of TikTok's position as a prominent social network for LGBTQ+ youth, as noted elsewhere (Ohlheiser, 2020). The proportion of disabled respondents is lower than census data for England and Wales (ONS, 2022): the audio-visual nature of the platform will have influenced this, but it is worth noting that TikTok has been shown to be biased against disabled users which may have discouraged use by disabled creators (Lyu et al., 2024; Köver and Reuter, 2019; Biddle et al., 2020). We were unable to establish if ethnicity data is reflective of the typical UK user base of TikTok, but it is relatively reflective of the UK population, per England and Wales census data (ONS, 2022). We asked respondents if they belonged to any other marginalised identities and responses included "working class", "neurodivergent" and "ex-sex worker".

6.4.1 All Respondents

We first present the results across all respondents, to paint a picture of typical use. To ensure our results across all demographics are indicative of the typical TikTok population, we use raking to weight the sexuality groups[13] per their prevalence in the *original* recruitment drive,[14] rather than after up-sampling of LGBTQ+ respondents. Typically, we report percentages, as is appropriate for weighted data, but where it improves clarity we also report raw n counts. We then analyse results across each demographic axis in turn, with a focus on how the experiences of marginalised users differ from non-marginalised users.

[12] Latest Scottish census data unavailable at time of writing

[13] The ANES raking variable selection algorithm (Pasek, 2018) determined changes to the sample based on gender and trans status were negligible

[14] For lack of more detailed TikTok user demographic information being available, we take our original sample distribution to be reasonably accurate

6.4.1.1 Use of TikTok

The vast majority (90.0%) reported having started using TikTok over a year ago. The majority (55.1%) of respondents used TikTok to view content over 3 times a day. The modal total time spent on the app was between 1-2 hours. This suggests our data is representative of "loyal" users of the app. The majority (74.1%) of respondents posted 0-3 times a month. This suggests most users of the app are relatively passive, consuming content but infrequently posting.

A very small number (3%) of respondents reported being ex-users of the platform. Of these, the most common reason for quitting was because content was no longer entertaining (reported by 55% of ex-users). "Too much moderation/ censorship" was reported by a single ex-user (raw count), suggesting this is not a primary motivating factor for leaving.

6.4.1.2 Experience of Censorship

The least common controversial content type that respondents posted was "Hate speech" (3.15%). The most common content types were "Curse words" (35.1% of respondents), followed by "Political content" (19.3%).

A relatively small percentage of respondents reported having had content removed (12.8%) or suppressed (14.1%). Around half of respondents who reported content suppression believed they had had content removed, and vice versa. Of those who report content removal, "Other" and "Content some may find offensive" account for the majority of reports of removal (reported by 38.1% and 33.5% respectively; respondents could select multiple content types). Of the "Other" content which had been removed, written answers included unwarranted concerns over minor safety (most common reason given), copyright infringement, drugs or alcohol, and content incorrectly identified as sexually inappropriate. That "Content some may find offensive or inappropriate" was the second most common suggests that respondents were aware (at least retroactively) that the content could offend. The most likely content type to "Always" be removed was violence.

For content suppression, the most common types of content that respondents believed had been suppressed were "Other" (26.0% of those who reported suppression) then "Curse words" (25.6%). Respondents who selected "Other" frequently reported being unsure why their content was being suppressed ($\sim 1/3$). Compared to content removal, reports of suppression were more evenly distributed across topics. This sug-

gests that the controversial content types are all considered possible reasons for suppression, even if removal is relatively uncommon for some e.g. Covid-related content. The most consistently suppressed content was "Sex related content for a non-erotic purpose".

The high number of fill in text answers suggests that the controversial topics we had identified through existing research explain only part of TikTok users' experiences of censorship on the platform (see *Limitations* in Section 6.6)

6.4.1.3 Beliefs about Algorithmic Censorship

Respondents most strongly agreed that "Content about violence that is intended to shock or disgust" is censored, selecting on average "Somewhat agree" (3.88). Respondents were least likely to agree that "Covid-related content" is censored, selecting "Neither agree nor disagree" on average (2.89).

We invited respondents to add anything else they believe about the TikTok moderation algorithm, and some themes emerged. Two respondents (raw counts) suggested the community guidelines were not being followed by the algorithmic censorship – for example, "They don't follow their own community guidelines because they ban innocent people all the time but keep people who break it for e.g. I always see nudity on my fyp". Many reported that content was removed without reason or that sometimes the "wrong" content was removed ($n = 12$). Some respondents ($n = 5$) shared that they felt the censorship was biased against marginalised groups, for example saying "Minorities are disproportionately affected... A black person can duet a racist person explaining that they are wrong, and the black person's video will be removed... not the racist's". Several respondents ($n = 3$) felt that moderation was politically motivated, for example because of its failure to remove propaganda.

6.4.1.4 Beliefs about Community Guidelines

There were several topics that users on average agreed were censored by the algorithm, but which they thought did not go against TikTok community guidelines (or were unsure). Of particular relevance to the focus of this paper, this was true for both content relating to minority identity experiences and content relating to a social justice movement. This was also true for non-erotic sex related content, curse words and political content. Respondents somewhat agreed (4.05) that when content was removed despite not violating community guidelines, this was due to other users reporting the content.

	Removal	Suppression
Male	-1.458 (0.768)	-0.469 (0.593)
Straight	-0.883^{*} (0.345)	-0.608^{*} (0.352)
White	–	-0.658^{*} (0.281)
Not Disabled	-0.850^{*} (0.331)	-1.004^{*} (0.326)
Male + Straight	$+2.223^{*}$ (0.826)	$+1.320^{*}$ (0.664)
Constant	-0.766^{*} (0.296)	-0.302 (0.389)

Table 6.2: Table showing coefficients and (standard errors) in two logistic regressions predicting content removal, and content suppression. $^{*}p < .05$.

This was the reason that respondents considered to be the most likely.

6.4.2 Experience of Censorship by Demographic Group

We now consider the impact of identity on experiences of censorship. We conducted logistic regressions to determine if identity predicts content removal and suppression. We exclude those who answered "Prefer not to say" to any demographic question ($n = 25$, 4.0% of data). We created dummy variables across gender, trans status, sexuality, disability and ethnicity (Marginalised by gender = 0, not marginalised by gender = 1, etc). We include binary interaction effects e.g. male + straight. We use a stepwise algorithm to determine the final model. We found disability status, gender and sexuality all impact content removal, $R^2 = .05, p < .001$: straight people and people without disabilities are less likely to experience removal, but straight men specifically are more likely to experience removal (see Table 6.2). We found disability status, ethnicity, gender and sexuality all predict content suppression, $R^2 = .06, p < .001$: white people, straight people and those without disabilities are less likely to experience suppression, but straight men are more likely (see Table 6.2).

Identity clearly impacts experiences of censorship, and it is not always the case that those of marginalised identities experience more censorship (i.e. straight men are more likely to report censorship). We now look at demographic attributes in turn to better understand experiences of censorship across identities.

Given the exploratory nature of this work, we did not formulate specific hypothe-

Figure 6.1: Chart showing the percentage of respondents by gender who posted "controversial" content, and reported having content removed or suppressed. Data labels show counts. Total counts are given after identity labels.

ses, but we do perform some post-hoc testing to suggest whether differences are substantive. We always exclude "Other" in our demographic comparisons due to small n and lack of homogeneity. We primarily conduct Fisher's Exact Tests to establish the significance of the difference in rates of posting controversial content or having it removed or suppressed, between groups. We report 2-sided p-values[15] and make suggestions as to whether the differences we report are likely to be meaningful. Where relevant we include other statistical analyses, noting that typically only large effect sizes can be detected. Tests were selected which were suitable to the lack of group size parity. This exploratory paper can indicate fruitful lines of future enquiry to reify our descriptive findings.

6.4.2.1 By Gender

Nonbinary respondents were the most likely to report posting one of the controversial topic types at least once, see Figure 6.1. Nonbinary people were more likely to post controversial content than men ($p = .050$) and women ($p = .001$). Nonbinary people were more likely to report having content removed and suppressed than men

[15] All p-values are from Fisher's Exact Tests unless specified

and women, see Figure 6.1; the differences between nonbinary people and women are significant ($p = .014$ for removal, $p = .016$ for suppression).

The types of content that respondents posted and reported as censored differed across genders. Men were significantly more likely to post political content than women (28.6%, $n = 64$, vs. 14.1%, $n = 53$; $p < .001$), and more likely to have it removed (11.76% of those who reported content removal, $n = 4$, compared to $n = 0$ women, $p = .019$), though a Chi-square test with history of posting political content as a layer (control) variable suggests this was entirely accounted for by history of posting. Despite these differences, respondents of different genders shared similar beliefs about censorship of political content. All groups neither agreed nor disagreed if it was subject to algorithmic censorship (3.15 for women, 3.09 for men, 3.25 for nonbinary people), and all disagreed that it went against community guidelines (2.45 for women, 2.29 for men, 2.15 for nonbinary people). Gender did not impact ratings, per Kruskal-Wallis tests.

Men were much more likely to post content some may find offensive compared to women (30.1%, $n = 68$ for men; 11.7%, $n = 44$ for women; $p < .001$), and more likely to report having this content removed (7.59%, $n = 17$ for men; 2.12, $n = 8$ for women; $p = .002$), though a Chi-square test suggests this was entirely accounted for by history of posting. Fifty percent of men who reported having content removed said it included content some may find offensive ($n = 17$). It was the most common type of content men reported having had removed. This was also the most commonly reported suppressed topic for men (28.6% of men who reported suppression, $n = 12$). Similarly, men were significantly more likely to post hate speech than women. Over 5% of men ($n = 14$) reported posting hate speech at least once (compared with 1.9% of women, $n = 7$; $p = .016$).

Nonbinary respondents were more likely than women to post content criticising a dominant group (20.1%, $n = 5$ vs. 6.37%, $n = 24$; $p = .022$), related to a social justice movement (30%, $n = 8$ vs. 15.6%, $n = 59$; $p = .042$) or about a minority experience (58%, $n = 5$ vs. 10.6%, $n = 40$; $p < .001$). Nonbinary respondents were also significantly more likely to post about minority experiences than men (vs. 16.1%, $n = 48$; $p < .001$).

Turning to beliefs about censorship, we find that men, women and nonbinary people differ in their beliefs about the censorship of content about marginalised identities (criticising a dominant group, related to a social justice movement or about a minority experience). We find that nonbinary people agree more strongly on aver-

age that this content is algorithmically censored compared to women, who in turn believe more strongly in algorithmic censorship than men. We conducted Kruskal-Wallis tests (which are suitable for small sample sizes) and included pairwise comparisons with Bonferroni correction, to minimise the risk of Type 1 errors. We find respondents did not differ significantly in their beliefs about the algorithmic censorship of content criticising dominant groups, but did so for content related to social justice movements ($H(2) = 19.909, p < .001$) and for content about minority experiences ($H(2) = 21.347, p < .001$). Nonbinary people were significantly more likely than men to agree social justice content ($p = .001$) and minority experience content ($p = .001$) were algorithmically censored; women were also significantly more likely than men to agree social justice content ($p = .001$) and minority experience content ($p = .032$) were algorithmically censored; nonbinary people were significantly more likely than women to agree minority experience content was censored ($p = .002$).

Our results suggest that there is not a clear relationship between history of censorship, and beliefs about algorithmic censorship. Nonbinary people did not differ significantly from men in terms of posting content related to a social justice movement, and there were no significant differences in reported removal or suppression for this topic (per Fisher's Exact Tests, $p = .209$ and $p = .186$ respectively), yet nonbinary people were more likely to agree this content is subject to algorithmic censorship. Likewise, beliefs about content suppression did not directly mirror content removal. For example, 14.3% of men who reported having content suppressed reported this content was related to Covid ($n = 6$), but less than 3% reported having Covid-related content removed ($n = 1$, 2.94% of those who report content removal).

Whilst respondents differed by gender in their beliefs about algorithmic censorship of content about marginalised identities, differences were not significant for beliefs about whether this content goes against community guidelines (per Kruskall-Wallis tests). Women, men and nonbinary people all disagreed that content related to a social justice movement or minority experience goes against community guidelines (for social justice content: 2.58 for women, 2.60 for men, 2.13 for nonbinary people; for minority experience content: 2.68 for women, 2.66 for men, 2.22 for nonbinary people). All three groups neither agreed nor disagreed if content criticising a dominant group goes against community guidelines (3.02 for women, 3.07 for men, 2.89 for nonbinary people).

6.4.2.2 By Trans Status

Comparing trans and non-trans respondents, we found that trans respondents were much more likely to post controversial content (75% of trans respondents, $n = 15$, vs. 50.6%, $n = 303$; $p = .040$). However, trans respondents were no more likely to have content removed or suppressed ($p = .163$ and $p = .185$).

Looking at data on the types of content posted, one significant difference is in the frequency of posting about minority identity experiences which is (somewhat unsurprisingly) much higher for trans compared to non-trans respondents (55%, $n = 11$ vs. 12.7%, $n = 76$; $p < .001$). Trans respondents were also comparatively more likely to report having this content removed (20% of those who report removal, $n = 1$ vs. 2.7%, $n = 6$; $p = .094$) or suppressed (60% of those who report suppression, $n = 3$ vs. 7.2%, $n = 6$; $p = .002$). Trans respondents also agreed more strongly that this content was subject to algorithmic censorship (3.79 vs. 3.31; $z = 2.436$,$p = .015$ per a Mann-Whitney U Test), although the two groups did not differ significantly in believing that this content does not go against community guidelines (2.40 for trans respondents and 2.67 for non-trans respondents).

There were further differences in the reported beliefs about censorship and community guidelines: for example, trans respondents more strongly agreed that "self-referential use of slur" was censored by the algorithm (3.96 vs. 3.44; Mann-Whitney U Test, $z = 2.385$; $p = .017$).

6.4.2.3 By Sexuality

Counts for posting controversial content and reports of removal and suppression suggest bisexual and asexual respondents were the most likely to report these experiences, as shown in Figure 6.2. To establish whether likelihood of posting controversial content differed significantly across sexualities, we first performed a cross-tabulation analysis with post-hoc z-tests across all sexualities to establish likely significant differences between sexualities. This suggested that bisexual respondents were more likely than straight respondents to post controversial content. We found this to be significant per a Fisher's Exact test (66.6%, $n = 65$ vs. 47.3%, $n = 218$; $p = .001$).

Repeating this procedure for content censorship, we find bisexual respondents were significantly more likely than straight respondents and gay respondents to report having had content removed. Fisher's Exact tests found these differences to be significant (22.2%, $n = 22$ for bisexual respondents vs. 11.5, $n = 53$; $p = .012$ for straight re-

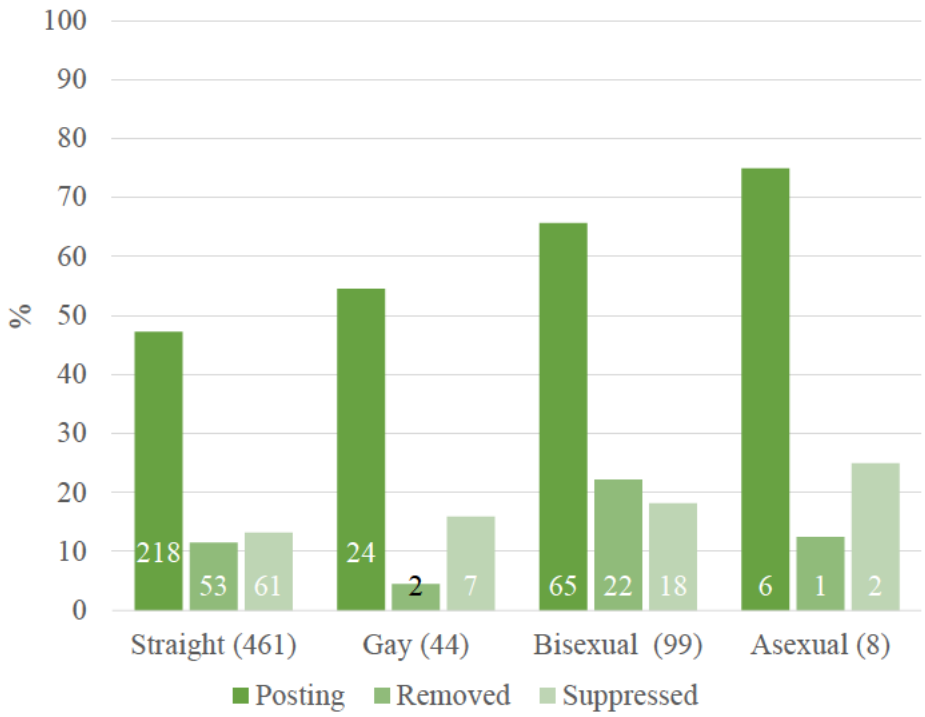


Figure 6.2: Chart showing percentage of respondents by sexuality who posted "controversial" content, and reported having content removed or suppressed. Data labels show counts. Total counts are given after identities.

spondents and; vs. 4.54, $n = 2$; $p = .008$ for gay respondents).

Whilst bisexual respondents greater posting of controversial content seems to explain the higher rates of censorship compared to straight respondents, this does not explain the comparative "lack of censorship" of gay respondents. Bisexual respondents report much higher rates of content removal compared to gay respondents, although a Chi-Square analysis with history of posting controversial content as a layer variable did not find this to be significant (for those with history of posting controversial content, $\chi^2(1,89) = 3.771, p = .084$). Given the small sample sizes, we cautiously suggest this test may have been "underpowered" and that this result merits further investigation.

As with other identity groups, we did not find a clear link between reports of censorship and beliefs about algorithmic censorship. For example bisexual respondents were more likely to agree that "sex related content for an erotic purpose" was subject to algorithmic censorship, compared to straight respondents (4.02 for bisexual vs. 3.70 for straight, per Mann-Whitney U test, $z = -2.349, p = .019$). However, there was no difference in posting rates between these two groups, nor did bisexual respondents report higher rates of censorship, suggesting this belief comes from observation of others' reports of censorship, or wider experience of the policing of queer sexualities, rather than direct experience of censorship on TikTok.

Straight respondents were most likely to report having content some may find offensive removed by the platform (43.4% of respondents who were straight and reported having content removed, $n = 23$), but this was reported by only two bisexual respondents (9.10% of bisexual respondents who reported content being removed). A relatively large number ($n = 6$) of straight respondents do *not* report posting content that some may find offensive, but *do* report having such content removed, perhaps indicative of a belief that the content they post is not offensive even if it is removed for being so. A comparison within those who report posting this type of content found bisexuals were less likely to have such content removed than straight respondents, but this difference was not significant (4.76%, $n = 1$ vs. 23.0%, $n = 17$; $p = .187$).

LGB+ respondents were much more likely to post about minority identity experiences – around $1/3$ of all gay, bisexual and asexual respondents report posting this type of content on average. Per a Fisher's Exact Test comparing straight and grouped non-straight (gay, bisexual, asexual) identities, this difference was significant, $p < .001$. This was also the most common type of content for bisexual respondents who reported content suppression: 33.3% ($n = 6$) of bisexuals who reported experiencing suppression, equivalent to $1/5$ of all who reported posting this type of content. This contrasts

Ethnic group	Posted	Removed	Suppressed
Asian	48.9% (22)	13.3% (6)	20.0% (9)
Black	66.7% (30)	11.1% (5)	20.0% (9)
Multiple	61.9% (13)	28.6% (6)	33.3% (7)
White	49.7% (254)	12.3% (63)	12.5% (64)

Table 6.3: Percentage of respondents in each ethnic group who reported posting controversial content, and having content removed or suppressed. Counts given in brackets.

with the fact that only two bisexuals reported having this content removed (9.10% of bisexuals who reported content removal), reflecting a disconnect between experiences of content removal and beliefs about content suppression.

6.4.2.4 By Ethnicity

Experiences of censorship differed by ethnic group (See Table 6.3). Black respondents reported posting the most controversial content, and were significantly more likely to do so than white respondents (66.7%, $n = 30$ vs. 49.7%, $n = 254$; $p = .030$), but no more likely than white respondents to report having content removed given a history of posting controversial content (per a Chi-Square analysis, $\chi^2(1,284) = 0.120, p = .812$, for those who post).

Almost a third of respondents of mixed or multiple ethnic groups reported content removal (28.6%) and suppression (33.3%); this is much higher compared to white respondents, even when controlling for history of posting controversial content: a Chi-square analysis found respondents from mixed or multiple ethnic groups were significantly more likely to report content suppression ($\chi^2(1,267) = 4.399, p = .046$, for those who post).

All marginalised (non-white) ethnicities report higher rates of content suppression compared to removal ($\sim 5\%$ greater or more) (see Table 6.3). For Black respondents, reports of content suppression were almost double those of content removal, whereas for white respondents levels were similar. People of colour may perceive themselves to face higher levels of censorship through suppression, even though censorship through removal is typically reported at similar rates across ethnicities (barring mixed ethnicity).

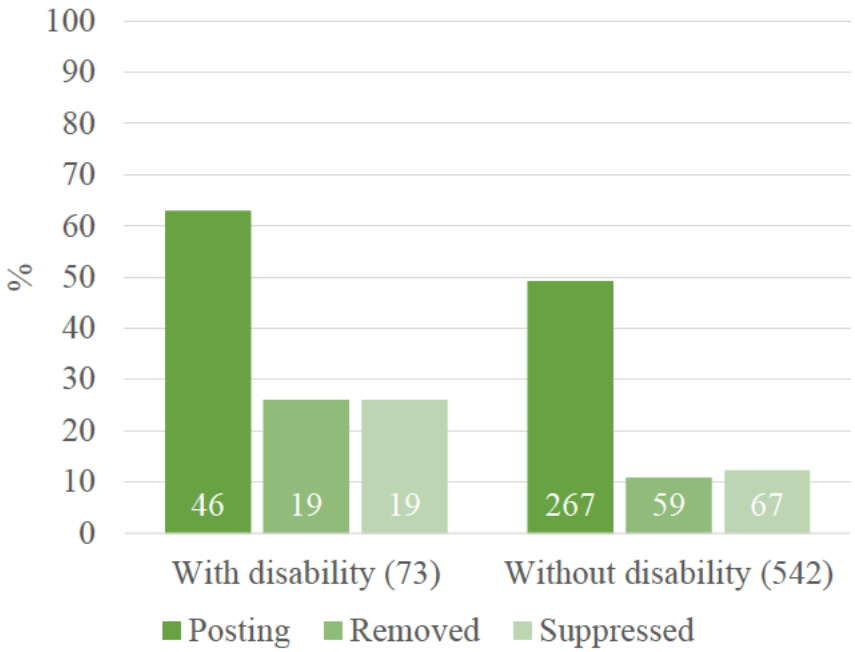


Figure 6.3: Chart showing percentage of respondents by disability status who posted "controversial" content, and reported having content removed or suppressed. Data labels show counts. Total counts given after identities.

6.4.2.5 By Disability

Respondents with disabilities were more likely to have content removed and suppressed, even when controlling for their greater likelihood to post controversial content (see Figure 6.3): Chi-square analyses with posting controversial content as a layer variable found respondents with disabilities to be significantly more likely to report content removal ($\chi^2(1,313) = 13.259, p < .001$) and suppression ($\chi^2(1,313) = 5.928, p = .022$).

Considering censorship of particular topics, as with LGBTQ+ minorities, respondents with disabilities were more likely to post about minority experiences compared to those without (28.8%, $n = 21$ vs. 12.2%, $n = 66$; $p < .001$). Respondents with disabilities were more likely to report this content being removed ($n = 3$ vs. $n = 0$). A Chi-square analysis with posting about minority experiences as a layer variable found respondents with disabilities to be more likely to report removal of this content ($\chi^2(1,87) = 9.765, p = .013$ for those who post). This may explain why respondents with disabilities seemed to more strongly agree that this content was subject to algorithmic censorship (3.58 vs. 3.30) (though this was not significant per a Mann-Whitney U Test,$z = 1.795$; $p = .088$) – although history of censorship did not always directly

relate to types on content posted. For example, our results suggest that respondents with disabilities may agree more strongly that self-referential slurs are censored by the platform (3.67 vs. 3.43) (though again this was not significant per a Mann-Whitney U Test, $z = 1.752$; $p = .080$) though there was no difference in rates of censorship.

6.5 Discussion and Future Directions

We identified a number of key trends in the data, which we discuss below. We also highlight avenues for future research to deepen our understanding of (perceived) biased censorship on TikTok.

6.5.1 Offence is in the Eye of the Beholder

We found that often content some may find offensive was one of the most commonly censored content types for non-marginalised groups (i.e. men and straight people). For example, straight people reported higher rates of removal than bisexual, gay or asexual people, and it was the most commonly removed content type; men reported higher rates of content removal than women and nonbinary people, and again it was the most commonly removed content type. Indeed, half of men who reported content being removed said it was content some may find offensive. This could relate to their greater tendency to post "Telling jokes" content compared to women and nonbinary people, which we also found, given the inherently subjective nature of humour. Meaney et al. (2022) explore gender differences in rating online humour and find that men "tolerate" offence more in "humorous" content. Further research into the nature of the potentially offensive content that men post is merited. It is also worth exploring the significantly higher rates of (self-reported) hate speech posting on the platform amongst men. It may be that men are no more likely to post hate speech, but are more likely to be honest about it. It may even be that men are less likely to take the topic seriously, so answered yes without much thought. Whatever the reason, it highlights the value of investigating hate speech content on TikTok, to understand who produces it and their attitudes towards offensive content.

As with conservative TikTok users compared to trans and Black users in Haimson et al. (2021), it may be that whilst straight and male respondents report being censored at comparable rates to marginalised respondents, they are having content removed that (they themselves agree) is against community guidelines, where marginalised creators

have content removed that does not clearly violate the platform's standards. Being censored for content that is believed to be in line with community guidelines will contribute towards feelings of discrimination and alienation. Users may feel they are being silenced for posting content that directly pertains to their lived experiences as marginalised people.

6.5.2 Direct Experience isn't Everything

We found beliefs about algorithmic censorship were not obviously correlated with direct experience of censorship on the platform. For example, nonbinary respondents agreed more strongly that social justice content is algorithmically censored, compared to men, even though nonbinary respondents never reported having this content removed. Thus, TikTok users' beliefs about censorship are not always the result of direct experience, but likely formed through which topics they (do not) see on the platform; the content they consume related to others' experiences of, and folk theories about, censorship; and censorship on other platforms and in "real life".

6.5.3 Suspicious Minds

We found there was a mismatch between reports of content removal, and content suppression. Around half of respondents who reported content suppression reported no content removal. Many more topics were reported as being subject to suppression compared to removal. Rates of suspected suppression were typically higher than rates of removal, across demographics. This suggests that beliefs about what content is suppressed by the platform are not directly related to experiences of content being removed (where a justification is typically provided). This is particularly evident when comparing across respondents from different ethnic groups. Reports of removal and suppression were at similar rates for white respondents, but respondents of colour were more likely to suspect suppression than removal. We do not attempt to say whether these suspicions are founded: it may be that users of colour are no more subject to suppression than white users, despite their beliefs. It may also be that TikTok favours suppression over removal of content by marginalised creators, to avoid similar scandals to Brown (2021); Ghaffary (2021); Ohlheiser (2021) *i.a.* (whilst suppression of marginalised creators' content has also received significant media attention (Kelion, 2019; Köver and Reuter, 2019), it is harder to prove than differing rates of content removal, account closures etc. – though cf. Ryan et al. (2020); Biddle et al. (2020).

Experimentation in the style of King et al. (2014) – who attempt to reverse engineer censorship in China – may help to shed light on which topics are censored through suppression on the platform. However, regardless of the "truth", that some marginalised users feel they are subject to additional suppression is harmful in and of itself.

6.5.4 Not So Innocent Mistakes

We found respondents felt censorship of content did not always align with the platform's community guidelines. Respondents reported that content was removed without reason or that sometimes the "wrong" content was removed. We found that for several of the controversial topics (namely non-erotic sex related content, curse words, political content, content relating to minority identity experience and content relating to a social justice movement), respondents agreed this content was subject to algorithmic censorship despite not being against community guidelines. However, whilst respondents acknowledged that algorithmic censorship could go against official community guidelines, when asked about their beliefs about why content might be removed or suppressed despite not violating guidelines, respondents most strongly agreed that this was due to other users reporting the content. This aligns with previous research which found human intervention was considered a primary cause for content removal (West, 2018). TikTok users have even expressed concerns about this being done to harass creators (Zeng and Kaye, 2022; Are, 2023).

6.5.5 The Bis Have It

We found the bisexual respondents may be subject to additional censorship compared to gay respondents, though we repeat our caution that this requires further empirical investigation. This additional censorship of bisexual content may be best understood in the context of the finding of Simpson and Semaan (2021) that the TikTok algorithm seems to favour queer content that aligns with norms. Bisexual people face what is known as "double discrimination", whereby they face rejection from both outside and within the LGBTQ+ community for their plurisexual identity (attraction to multiple genders) (Mereish et al., 2017). It is possible that content related to bisexual experiences falls outside the norm that has been constructed on TikTok – by its creators and users – hence bisexual respondents were subject to greater levels of censorship than gay respondents.

6.6 Limitations

We did not format our questions around TikTok's specific moderation guidelines (e.g. related to copyright music), which partly explains the high number of "Other" answers to types of censored content. This may have biased our findings against topics of particular relevance to TikTok such as minor safety. Our sampling method meant we obtained limited data from individuals with direct experience of censorship. An alternative method would be to target creators on TikTok who report censorship, or create videos to promote the study, but in both cases algorithmic curation would introduce a sampling bias. Our use of the term "censorship" may have influenced responses, as censorship has more negative connotations than for example "moderation". Future work might employ more neutral language i.e. only using the terms "removal" and "non-recommendation". We asked which types of content people post and which types are censored as separate questions, and some respondents gave conflicting answers. Mismatches may have arisen due to respondents preferring to select "Never" for the frequency of posting a type of content they have posted only once. Or respondents may have interpreted the content type descriptions differently in each instance, e.g. they do not believe the content they post to be sexual, but this was the reason given for its removal. Greater clarity in our instructions would have addressed this issue.

Our sample may not be representative and we vehemently discourage use of the work to conclude that specific groups do not experience harms from censorship, simply because we did not find evidence.

6.7 Conclusion

Our exploratory analysis has revealed that whilst rates of censorship are not always higher for marginalised creators, they are more likely to be censored for content that is generally regarded as in line with community guidelines, through removal and suppression. This will contribute towards feelings of discrimination and alienation on the platform. We have highlighted several avenues for future work into experiences of biased censorship on TikTok, to complement investigative work already conducted by journalists. We encourage a focus on users' beliefs over any attempt to identify the "truth", as these folk theories of censorship play a pivotal role in shaping users' experiences of the platform.

6.8 Learnings

This work highlights the value of asking those impacted by AI about their beliefs and not just direct experiences. For example, if we only recorded rates of content removal we might assume that this determines what content people believe is censored – but actually beliefs about censorship clearly stem from much more than direct experience. We would also fail to capture beliefs about rates of suppression, which for example are higher amongst people of colour than white people. This is valuable because it is these beliefs that will guide people's behaviour on the platform, and their emotional response. Where people are being harmed, they are being harmed not just by the actual output of the model but also by their beliefs about the model – their beliefs about what it represents. As Ehsan et al. (2022) writes, negative interactions with algorithms change how people make sense of the world.

This work also highlights the importance of not treating the LGBTQ+ community as a homogeneous group – either with regards to stimuli or respondents – because once again I found differences within the queer community. For example, I found that bisexual people seem to experience unique discrimination compared to monosexual queer identities (i.e. gay and lesbian). Results that offer an average over all queer identities may fail to capture important differences (e.g. bisexual people often face "double discrimination" from the queer and non-queer community, which I return to in Appendix A).

This work exemplifies the other maxims of my proposed human-centric approach to bias research, in that I consider the human and technological aspects of the moderation and recommender systems, for example asking respondents about the possible role of human moderators, other users and "the algorithm" in their experiences of biased censorship.

Chapter 7

Le$bean or Lesbian? A Survey of Marginalised Users' Motivations for Obfuscation on TikTok

Having established beliefs about biased censorship on TikTok, I now explore the linguistic techniques that platform users employ to obfuscate – hide – their intended meaning, in order to evade perceived censorship. This has received significant media attention, and in this Chapter I complement this media interest by conducting a survey to establish users' motivations for employing these techniques, and their attitudes towards this behaviour. This work is informed by linguistic scholarship on self-censorship, anti-languages and platform vernaculars. I conducted a survey of 627 UK TikTok users and found that use of obfuscation was relatively low in our sample, and primarily related to the types of content users were posting (historically censored content), rather than acting as a way to establish social identity as I had predicted – though our sample was far from homogeneous on this point. Using a structural equation modelling (SEM) analysis I show that men and people of colour (POC) were significantly more likely to use obfuscation. For POC this is driven partly by positive associations with obfuscation use, suggesting obfuscation is seen as a way to be playful and creative with language, as well as evade "The Algorithm". This makes particularly visible the work that the public do to understand the role that biased AI plays in their lives: as the output of AI impacts them, so too do they impact the output of AI. Studying this reciprocal relationship is key to a human-centric approach to studying bias.

7.1 Introduction

The short-form video content app TikTok's meteoric rise (Harwell, 2022) has not been without controversy: users complain of experiencing unfair censorship on the app, in particular those who are already marginalised, for example by their gender, ethnicity, sexuality or disability, as we discuss in the previous Chapter (see also Ryan et al. (2020); Brown (2021); Karizat et al. (2021); Lyu et al. (2024)). As a result, users will often use homonyms, misspellings, leet speak (Blashki and Nichol, 2005) and other obfuscation techniques to hide their intended meaning from "the algorithm" (Calhoun and Fawcett, 2023, 2022; Klug et al., 2021). For example, users might talk of being a member of the "leg booty" community, rather than LGBT+, out of concern that their video will be suppressed (Lorenz, 2022). This subject has garnered extensive media attention (Lorenz, 2022; Brown, 2021; Bacchi, 2020; Ohlheiser, 2021). Calhoun and Fawcett (2022, 2023) provide a qualitative analysis of the most prominent obfuscation techniques observed on the platform, and propose likely motivations for use such as establishing community and avoiding censorship. Building on this, we conduct a large-scale survey investigating users' motivations and attitudes. Our quantitative analyses include measuring the relative importance of different motivations in determining obfuscation use, through an SEM, well-suited to studying causal relationships between many constructs (Hair et al., 2019). This work explores how folk understandings of "the algorithm" (see Section 7.2) affect user behaviour on TikTok (Klug et al., 2021), and in doing so demonstrates the value in considering users' beliefs when studying their interactions with information technologies.

Crucially, our focus is not on whether, or how censorship occurs, or whether these obfuscation techniques are truly successful at evading it. We look to address users' perception of being censored, and their motivations for using, and attitudes towards, obfuscation techniques,contributing to the body of work addressing users' beliefs about digital technologies including social media algorithms (Simpson and Semaan, 2021; Klug et al., 2021; Lyu et al., 2024, i.a.), but also smart speakers (Meng et al., 2021; Frick et al., 2021) and algorithmic decision making systems (Starke et al., 2022). Though we do not investigate how the censorship system "really works", we do (in line with Karizat et al. (2021)) acknowledge that this user behaviour may have very real effects on how the moderation algorithm functions, in that users' attempts to resist an algorithm they perceive to be biased undermines how effective the algorithm can be. Social media companies need to understand and anticipate user behaviours in order to

ensure their moderation systems have the desired effect. As Seaver (2019, 418) notes it is "intricate, dynamic arrangements of people and code" which have "sociocultural effects", not algorithms in isolation (see Section 7.2.2).

In addition to avoiding censorship, creative language use also plays a role in establishing online group membership (Stewart et al., 2017; Blashki and Nichol, 2005), as well as preventing out-group readers from understanding (Halliday, 1976). We will also investigate to what extent establishing membership of a community motivates users to employ obfuscation techniques. This could apply both to feeling part of a particular cultural identity, for example the LGBTQ+ community, but also to feeling part of the wider TikTok community, as it could be said use of obfuscation techniques has become part of the platform's culture.

Whilst creative use of language to avoid censorship is for many users a way of enabling positive and informative content to be viewed by others (Lorenz, 2022), if these obfuscation techniques are successful this may also allow inappropriate content to circulate on the platform. Automated content moderation is designed to prevent the latter, but this can come at the expense of suppressing otherwise "innocent" content (Haimson et al., 2021). We investigate the kinds of content users post when using these techniques, to establish how frequently obfuscation is used to post inappropriate content, and if this is a primary motivation.

We conduct a survey of active and historic TikTok users, to understand their motivations for using obfuscation techniques, be that to avoid censorship or to signal in group identity. We also study what how feel about this practice, as we particularly wish to establish whether users feel harmed by the need to adapt their use of language to avoid censorship of their content. We are particularly interested in investigating the motivations and consequences of these strategies for already marginalised groups such as Black and LGBTQ+ users of TikTok, groups which media coverage suggests feel most driven towards obfuscation use by perceived censorship (Lorenz, 2022; Brown, 2021; Bacchi, 2020; Ohlheiser, 2021). A primary contribution is to demonstrate how motivations for use differ between demographic groups.

In the following, we provide a brief introduction to TikTok, then present a literature review which informed the development of our hypotheses and thus of our survey. We present our methodology, then analyse and discuss our findings. Finally we discuss the implications of our findings, for those researching social media, but also for those running these platforms. We put particular emphasis on the implications of our findings with regards to how social media platforms should engage with users around the

moderation of content on the platform in order that minority communities do not feel harmed by these practices.

7.2 Background

7.2.1 Overview of Censorship on TikTok

Moderation and recommender systems allow social media platforms to create a safer online environment, as they enable the management of harmful content at scale. This may be necessary to comply with legal requirements, such as the UK's Online Safety Act (DSIT, 2025), to align with the platforms' stated agendas such as "amplifying LGBTQIA+ voices",[1] and to meet marginalised users' expectations around safety on the platform. Content moderation is intended to protect users, but also acts to control and censor (Cobbe, 2021; Zeng and Kaye, 2022). In this Chapter, we again use censorship to refer to both removal and suppression of content, including of content that would not intentionally be subject to moderation.

Whilst social media censorship is common, TikTok has garnered substantial media attention (Hern, 2019a,b; Lorenz, 2022; Brown, 2021; Bacchi, 2020; Ohlheiser, 2021; Biddle et al., 2020; Woods, 2021; Kelion, 2019), likely due to inconsistent policies, its association with China, and the central importance of algorithmically curated content. Although TikTok tailors content moderation to local cultural norms (Ryan et al., 2020; Hern, 2019b), such moderation is often evident to users globally as it operates at the level of language, meaning it is more open to critique. Further, TikTok has previously been criticised for promoting Chinese foreign policy aims (Hern, 2019a). Finally, user awareness of algorithmic censorship may be particularly high because the platform's design encourages viewing of algorithmically recommended content over for example followed accounts.

7.2.2 "The Algorithm" and Algorithmic Folk Theories

In this Chapter, we follow Seaver (2019) in conceptualising algorithms as socially constructed objects, rather than adopting a more narrow definition from computer science. A socio-technical lens allows us to understand how social media users construct "the algorithm" as a mediating force affecting which content they are shown (and con-

[1] https://newsroom.tiktok.com/en-us/celebrating-the-lgbtqia-community-on-tiktok-and-beyond

versely, which content they are not shown). In line with this social construction, we refer to "the algorithm" throughout this Chapter, in the understanding that algorithmic content moderation systems are both technically and socially more complex.

TikTok users feel that content is suppressed by the platform's "algorithm" on the basis of the creator's identity (Karizat et al., 2021) - an example of a algorithmic folk theory, which shapes user behaviour on social media (West, 2018; Karizat et al., 2021). Such folk theories often develop from "everyday algorithm auditing": the discovery of problematic behaviour through every day interactions with a system (Shen et al., 2021; Karizat et al., 2021; Simpson and Semaan, 2021; Brown, 2021). A detailed discussion of these topics can be found in Chapter 6..

Akin to these everyday audits, journalists at The Washington Post have created videos to test whether certain terms are suppressed.[2] They leave out terms they believe may trigger censorship, including the word "suppressed" which they display in the video on a handheld sign (arguably a form of obfuscation). This kind of work supports that done by everyday users, and as Velkova and Kaun (2021) note, algorithmic resistance work may become "increasingly dependent on collaboration with traditional media".

7.2.3 Algorithmic Resistance

To counteract perceived content suppression and the resulting "algorithmic representational harms" (Karizat et al., 2021), users may engage in forms of "everyday algorithmic resistance" whereby they attempt to change how the platform behaves whilst operating within its framework. An example of this is deliberately engaging with certain content to ensure it is not suppressed by TikTok (Karizat et al., 2021; Simpson and Semaan, 2021), a technique that is reminiscent of data poisoning attacks, that is, the manipulation of training data to degrade the accuracy of a machine learning model (Pang et al., 2021). Velkova and Kaun (2021) write that just as algorithms study users, so too do users study algorithms. This may be undertaken as an individual or collective action. Understanding these "tactics of resistance" (Velkova and Kaun, 2021), which allow users to express their agency, is key to developing approaches to manage the power that algorithms have in everyday life (Velkova and Kaun, 2021).

The term everyday algorithmic resistance could also be taken to encompass the use of coded or obfuscated language online to avoid automated censorship. Coded lan-

[2] https://www.tiktok.com/@washingtonpost/video/7158098211977088298

guage, including visual language such as emoji and memes, can allow internet users to express dissatisfaction with their government and with strict censorship (Mina, 2014), or to avoid moderation of sensitive content (Calhoun and Fawcett, 2022, 2023). Obfuscation is one of the methods through which users can redefine the "functioning [...] of technologies, collectively and politically" (Velkova and Kaun, 2021). Automated content moderation is not new, nor is use of obfuscation to avoid it: we found reference to use of alternating case to avoid profanity filters in a 1998 edition of PC Mag (Lidsky, 1998).

Such coded language is also used by those wishing to promote misinformation (Collins and Zadrozny, 2021) and spread toxic content (Lees et al., 2021). These latter two behaviours could also be said to constitute resistance to automated censorship. However in this Chapter we choose to focus on algorithmic resistance by marginalised communities, as we wish the understand the algorithmic representational and allocational harms that may be done to these communities through use of automated censorship, in line with the focus of this thesis.

7.2.4 Obfuscation Technique Use on TikTok

A prominent form of algorithmic resistance is the use of obfuscated language to avoid censorship. For example, users might write "yt" instead of white, "le$bean" instead of lesbian, or "shrek work" instead of sex work (Calhoun and Fawcett, 2022). This creative use of orthography is designed to evade the automated censors whilst still being comprehensible to (human) viewers. Use of obfuscated text has been found on Facebook (Collins and Zadrozny, 2021) and Instagram (Chancellor et al., 2016; Stewart et al., 2017), and its use on TikTok has received much media attention (Lorenz, 2022; Woods, 2021; Tait, 2022; Harwell, 2022). We have also casually observed obfuscation technique trends spreading from TikTok on to other platforms, just as the video format has spread (Harwell, 2022; Kantrowitz, 2023; Frier, 2022). State-of-the-art obscene language detection systems show success at detecting obfuscated content that use spaces, punctuation and letter substitution (Renwick and Barbosa, 2021) (cf. Röttger et al. (2021)), and TikTok may employ similar models, though it is not unclear from their official documentation[3] what systems they use to moderate content, nor how responsive it is to users' changing behaviours.

[3] https://www.tiktok.com/transparency/en-us/content-moderation/

Calhoun and Fawcett (2022, 2023) identify eight key obfuscation techniques used in "linguistic self-censorship" on TikTok. As they point out, avoiding automatic censorship is not the only reason for this linguistic behaviour: in addition to enabling users to discuss sensitive topics, it is a playful way to express linguistic creativity, (social) identity and stance (Calhoun and Fawcett, 2023). These social and creative motivations account for the use of obfuscation techniques in the context of topics not deemed sensitive or controversial according to platform guidelines (Calhoun and Fawcett, 2023). Building on these initial studies, we further explore (minority) users' social motivations for using obfuscation techniques.

7.2.5 Coded Language, Secrecy and Social Meaning

Of course, use of "obscured" language to avoid detection by out-group members (and ensure visibility to in-group members) is not new amongst marginalised communities. Halliday (1976) introduces the term "anti-language" to describe "a language generated by a kind of anti-society" or groups positioned outwith mainstream society, which Calhoun and Fawcett (2023) write exist at one end of a spectrum of language play. These "anti-languages" might differ from a language spoken by the mainstream society only in (parts of the) lexicon, particular in domains that are central to the subculture. As Halliday (1976) points out, in this way anti-languages are similar to specialised registers used by particular social groups which are extremely common and well-studied both in mainstream society and subcultures (Bucholtz, 1999; Bieswanger, 2016; Tway, 1975). What sets an anti-language apart, on a linguistic level, is a very high number of synonyms in particular domains. Playful and original language use and secrecy as motivations for "overlexification" and, importantly, the inherent counter-cultural stance of its speakers, is what sets it apart on a social level. Crucially, the secrecy (or obfuscation) is not the main reason why these languages emerge, but rather they are used by the speakers to create and maintain their "anti-societies" or subcultures. In addition to the examples related to "criminal underworlds" that Halliday (1976) provides, Polari, a linguistic variety predominantly used within gay male subcultures in the twentieth century, can be understood through this lens (Baker, 2019).

In the context of social media, characteristic uses of particular linguistic and aesthetic repertoires have been understood as "platform vernaculars" (Gibbs et al., 2014). Aspects of these vernaculars may well be used to express affiliation with particular online or offline subcultures (Calhoun and Fawcett, 2023). The adoption of obfusca-

tion techniques, like the adoption of anti-languages, might also be motivated by (and expressive of) a kind of resistant stance towards (algorithmic) censorship as described by Velkova and Kaun (2021).

7.3 Hypothesis Development

In this study we seek to understand marginalised users' motivations for using obfuscation techniques. We will explore the kinds of content users are posting when they use obfuscation techniques; how technique use differs across demographics and across those with a history of content removal. Accordingly, our research question is as follows: (R1): How do motivations for use of obfuscation techniques differ between identities? In addition, we make a number of specific predictions about minority users' motivations for obfuscation technique use.

Obfuscation use on TikTok has received some media attention (Lorenz, 2022; Woods, 2021; Tait, 2022; Harwell, 2022), with articles typically sharing the intuition that usage relates to avoiding algorithmic censorship. Calhoun and Fawcett (2022, 2023) write that users' motivations also include the desire to express linguistic creativity. We argue that like anti-languages, obfuscation can be used to index group affiliation and remain illegible to "outsiders" (understood here as both automated and human moderation, and those who might report their content). We predict that, reflecting a history of anti-language use by marginalised groups (Halliday, 1976; Baker, 2019):

H1A: Minority groups will be more likely to use obfuscation techniques than non-minority groups

Media coverage of and articles on TikTok censorship suggests increased censorship of marginalised users (Brown, 2021; Lorenz, 2022; Kelion, 2019; Ohlheiser, 2021; Köver and Reuter, 2019; Haimson et al., 2021; Karizat et al., 2021; Simpson and Semaan, 2021) – we found evidence of this in Chapter 6. We wish to understand the role perceived censorship plays in motivating use of obfuscation techniques. We predict that, in line with media coverage, that:

H1B: Minority users will (perceive themselves to) face additional censorship

As avoiding perceived censorship is one of the motivations for use of obfuscation (Calhoun and Fawcett, 2022, 2023; Lorenz, 2022), we predict that:

H1C: Higher perceived censorship will predict greater use of obfuscation techniques

In addition to avoiding censorship, obfuscation use can play a number of sociolinguistic roles as explored above, and by Calhoun and Fawcett (2022, 2023) (mirroring the role of anti-languages such as Polari (Halliday, 1976; Baker, 2019)) including rendering text illegible to out-group members. That is, obfuscation technique use may be unrelated to censorship. Thus we predict that:

H1D: Minority users are motivated to use obfuscation techniques to maintain a social identity

Namely, we will investigate the impact that social motivations (emotional responses to obfuscation use, sense of community membership) have on obfuscation technique use.

To deepen our understanding of users' responses to censorship on TikTok, we also investigate users' beliefs about algorithmic censorship and obfuscation techniques. That is to say, we seek to contribute to our understanding of users' folk theories about censorship on TikTok (Karizat et al., 2021). We will explore what content users believe to be censored, and how users interpret censorship "mistakes" (when content is censored that follows community guidelines). Additionally, we predict that (given their use), users will believe obfuscation techniques to be effective, but that low belief in effectiveness will discourage use of obfuscation techniques, moderating the impact of a history of posting controversial content, and of perceived censorship by the platform (i.e. those factors related to use of obfuscation techniques to avoid censorship).

H2: Perceived effectiveness of the obfuscation technique moderates the effect of perceived censorship and controversial content posting on the use of the obfuscation technique

7.4 Methodology

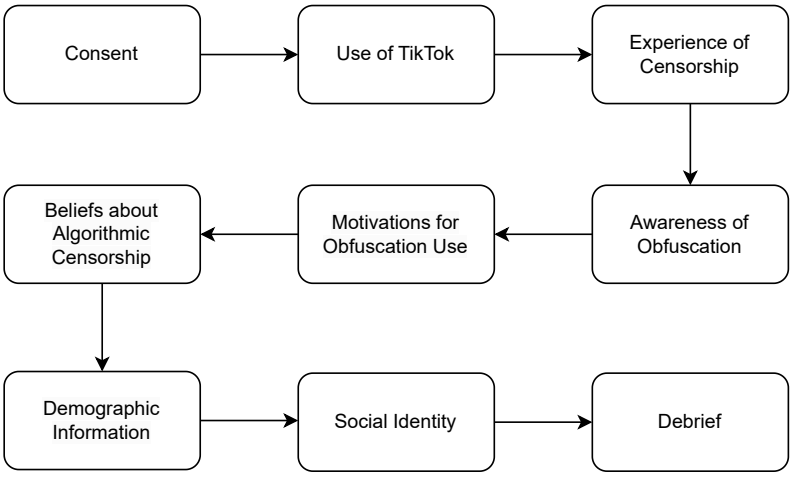
7.4.1 Recruitment

The recruitment of our respondent pool is discussed in Chapter 6.

7.4.2 Procedure and Measurements

We conducted a survey. Items in the survey were used to measure latent variables in our structural equation model, plus the observed variable of obfuscation use in our logistic regression. We additionally conduct post-hoc analysis of differences across

Figure 7.1: Flowchart showing sections of survey



user groups. Ethics approval was obtained from the University of Edinburgh Informatics Research Ethics Process, rt #6862. This study was conducted online using Qualtrics.com.[4] Qualtrics follows best practice for data security, vital given our focus on sensitive demographic information and posting of potentially controversial content. Before agreeing to the survey, respondents were given a detailed information sheet, which specified that data would be kept strictly confidential and viewed only by the named research team. They were then asked for their consent.

7.4.2.1 Use of TikTok

After giving informed consent, respondents were asked about their use of TikTok, and their motivations for stopping use if they reported being ex-users. Respondents were asked how often they used TikTok to view, and separately to post content. We asked about the types of content users viewed and separately posted on TikTok. We told users that for this survey we were interested in their use of TikTok for English language content only, because our hypotheses had been developed based on research into English language users, and the obfuscation examples we gave were all English language based. Further details can be found in Chapter 6.

7.4.2.2 Experience of Censorship

We asked about users' experience of censorship. We stated that by censorship "we mean both when content is removed and when content is suppressed." We gave the ex-

[4] Qualtrics.com

ample of a video getting very few views as something that might indicate suppression, which is reflective of TikTok users' folk theories (Lyu et al., 2024).

We provided a list of 13 topics (henceforth "controversial topics", referring to the fact that users have previously reported having such content removed) derived from Haimson et al. (2021), and the option to supply "Other". The list (repeated from Chapter 5) is as follows:

- Political content
- Content some may find offensive or inappropriate
- Sex related content for a non-erotic purpose i.e. that is intended to educate
- Sex related content for an erotic purpose
- Covid-related content
- Content insulting or criticizing dominant group (e.g., men, white people)
- Content relating to a social justice movement, for example feminism or anti-racism
- Content relating to minority identity experience i.e. queer content, content about Black experiences
- Hate speech
- Curse words
- Self-referential use of slur i.e. d*ke by a lesbian
- Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime
- Content about violence that is intended to shock or disgust

By providing a pre-defined list we ensured we had consistent data across respondents. Users could indicate whether this kind of content had been removed on a 5-point scale of "Never" to "Always" (such frequency scales are standard in psychological research; 5-point scales show good reliability and ease of use (Preston and Colman, 2000), and we preference 5-point scales throughout given ease of use on mobile screens). We repeated this process for content suppression on TikTok. We then asked questions irrelevant to this Chapter.

7.4.2.3 Awareness of Obfuscation

Next, respondents were introduced to the concept of "obfuscation" with examples shown in screenshots taken from TikTok. We gave explanations and multiple text examples of the eight key obfuscation techniques identified by Calhoun and Fawcett

(2022) and asked which of these techniques users had seen others using. For those who indicated that they had posted content, we also asked which they had used themselves. We asked how often respondents understood the intended meaning when they used these techniques, with five options from "Never" to "Always" (1-5). We asked respondents to rate how effective these techniques were for avoiding algorithmic censorship, on a 5-point scale from "Not effective at all" to "Extremely effective" (1-5). Users were given the chance to add any other thoughts on the topic.

Users were then asked how using or seeing obfuscation techniques made them feel using the 20 item mDES scale (Fredrickson et al., 2003) (also adapted in Fredrickson (2013)), on a 5-point agreement scale from "strongly disagree" to "strongly agree" (original scale looked at frequency of emotion experienced over the last two weeks - this scale was inappropriate given our focus). $\alpha = .937$ for positive emotions in response to obfuscation techniques, $\alpha = .938$ for negative emotions.

7.4.2.4 Motivations for Obfuscation Use

We then asked respondents which kinds of content they had posted with five options from "Never" to "Always (all my content is about this)", repeating the 13 item controversial topics list. For topics they selected, we asked how often they posted *using obfuscation techniques* with five options from "Never" to "Always".

In order to explore other motivations for obfuscation technique use beyond post topic, we asked respondents to state how strongly they agreed with the following statements, on a scale of 1-5 ("Strongly disagree" to "Strongly agree"): "To express my creativity", "Because it's fun", "So human moderators can't understand what I am writing", "So the algorithm does not censor my posts", "So only certain TikTok users can understand me", "To show which communities I belong to", "To feel part of the community of TikTok users", "To protest algorithmic censorship". These eight original items, used in our exploratory descriptive analysis, were devised to measure users' explicit motivations for using obfuscation. Items 1 & 2 were taken to indicate creative language use motivations for obfuscation technique use the online "language play" explored in Calhoun and Fawcett (2022, 2023)). Items 3 & 4 relate to use of obfuscation to avoid moderation or censorship. Items 5-7 were taken to indicate community establishing motivations for obfuscation technique use. Item 8 was intended to measure whether respondents who were consciously aware of obfuscation use being a form of algorithmic resistance saw it as an act of protest.

7.4.2.5 Beliefs about Algorithmic Censorship

Users were asked if they agreed that the moderation algorithm censors at least some posts about the 13 controversial topics in turn, on a scale of 1-5 ("Strongly disagree" to "Strongly agree"). We then asked questions irrelevant to this Chapter.

7.4.2.6 Demographic Information

Respondents were asked to optionally provide demographic information, namely age, gender identity (male, female, non-binary, other [text entry]), sexuality (straight, gay, bisexual, asexual, other [text entry]), disability status (yes, no), ethnic group (topline categories taken from the UK 2021 census [5]). We then asked questions irrelevant to this Chapter.

7.4.2.7 Social Identity

Finally, we asked respondents how strongly they agreed to five statements relating to their social identity community and TikTok. We provided a definition of social identity as meaning belonging to a certain social group, "such as a certain race, gender or class" taken from Karizat et al. (2021). We asked "on TikTok..." "I feel like I am part of my communities", "I feel a connection with other members of my communities", "I feel accepted by other members of my community", "I feel respected by other members of my communities" and "I feel valued by other members of my communities". These questions were adapted from existing work on sense of belonging in academic communities. Namely we used a subset of the belonging measures found in Muradoglu et al. (2021), leaving out three statements which pertained to emotions felt when "around" the community (Muradoglu et al. (2021) was in turn adapted from Good et al. (2012)).

Respondents were then debriefed.

7.5 Structural Equation Model

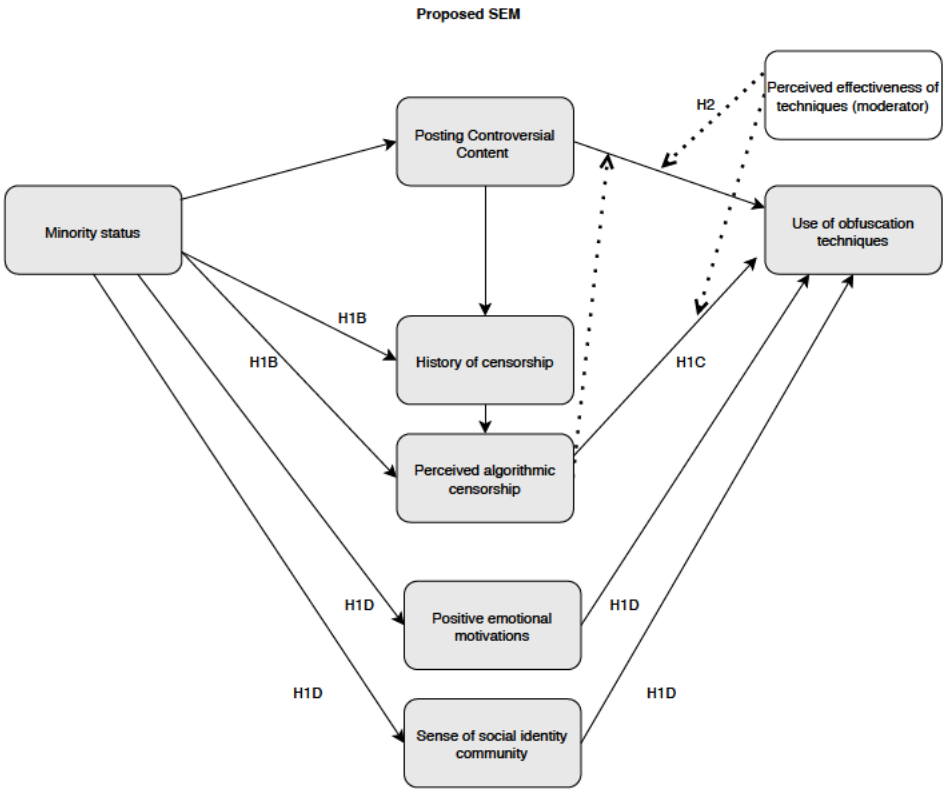
We first present the results of a structural equation model looking at marginalised respondents' motivations for using obfuscation. The model attempts to establish which factors including posting controversial content influence obfuscation use. Analysis of the text results show very, very few respondents reported using obfuscation techniques for any topics other than those we had pre-identified ($n = 5$), so we do not consider

[5] https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups

it a detriment of the model that it focuses on the pre-determined controversial topics. Because we are trying to predict a binary variable (whether the respondent used obfuscation or not) we run our analysis in two stages, per best practice proposed by Bodoff and Ho (2016). Specifically, we first run a partial least squares (PLS)-SEM algorithm to establish latent variables and path coefficients for endogenous variables that are not binary (i.e. other than "use of obfuscation"). We then run a logistic regression using the latent variables output by the PLS-SEM model to predict "use of obfuscation techniques".

7.5.1 Design

Figure 7.2: Proposed models relating to hypotheses for H1A-D & H2. Arrowheads indicate a direct effect. Dots indicate a moderation effect.



To test whether factors other than perceived censorship and content topic influenced marginalised respondents use of obfuscation techniques, we designed an SEM, depicted in Figure 7.2. PLS-SEM is suited to modelling data with a view to provide

a causal explanation, and where the proposed model has many constructs (Hair et al., 2019). Conducting many individual regressions could compound the risk of Type 1 errors. To conduct our analysis we used an academic license for SmartPLS 4 (Ringle et al., 2022). We used a path weighting approach in our analysis, the recommended approach for PLS-SEM (as per SmartPLS documentation[6]). Path weighting is based on regression rather than just correlation between latent variables (Henseler et al., 2009). The initial weight of the relationship between observed and latent variables was set by default to +1. Significance values were determined using bootstrapping, with 5000 subsamples and two-tailed significance determined at $p < .05$.

We removed respondents who did not provide the complete set of demographic information ($n = 25$).

Frequency of posting controversial content, content removal and suppression were all represented originally by ordinal data (0-4 to represent "Never" to "Always"). This data was taken as quasi-continuous and thus appropriate for SEM.

The observed variable (see Bodoff and Ho (2016) for a discussion of this distinction) of "use of obfuscation techniques" was recorded as whether respondents selected "None" for obfuscation use.

Minority status was originally captured by five latent variables, each with a single binary dummy observed variable recording whether the respondent belonged to a marginalised group across gender,[7] sexuality, trans status, ethnicity and disability (binary measurement variables are permissible in exogenous variables in PLS-SEM). We anticipated these demographic groups to exhibit similar patterns with regards to obfuscation use to those reported by historically marginalised users in Haimson et al. (2021). In the first version of the model, we saw no significant relationships between trans status and any connected latent variable in the model, despite existing evidence to suggest a relationship between trans status and controversial content posting (Haimson et al., 2021), likely due to the very small number of trans respondents ($n = 20$). For simplicity's sake, we removed trans status from the model before reporting the final findings.

Note, an alternative way of modelling minority users' behaviour would be to exclude any respondents from the analysis who do not face marginalisation along any

[6] https://www.smartpls.com/documentation/algorithms-and-techniques/pls/

[7] Note that due to the way we conceptualised marginalisation due to gender, the five trans men in our sample were included in the non-marginalised group for gender (on the basis of their identity as men). It could be argued that their trans history makes them likely to experience marginalisation due to their historical or perceived gender, but we chose to model gender and trans status separately.

of our identified axes of oppression (gender, sexuality, trans status, ethnicity and disability), and thus analyse the behaviour of a single "marginalised" group. However, this design of model would not allow us to model the impact of marginalisation along a particular axis of oppression on any of the predictors of obfuscation use. Using a single latent variable to represent all forms of marginalisation would be inappropriate given the binary measurement variables, which cannot be combined into a single formative latent variable (and further, would not allow us to explore substantive differences between the different marginalised groups). Creating separate models for each demographic would have increased our risk of Type 1 errors.

We predicted minority respondents would make greater use of obfuscation techniques (H1A), in part by virtue of the following four mediating latent variables: perceived censorship (of controversial topics), controversial content posting, positive emotions associated with obfuscation technique use and social identity community motivations.

We predicted minority status would lead to greater perceived algorithmic censorship (H1B) (which is also influenced by minority status by way of history of content posting and of censorship). Perceived censorship was originally formed of ratings for perceived censorship for all the controversial content types, but we removed the majority due to non-significant weights and loadings (per PLS-SEM best practice (Hair et al., 2019)), leaving perceived censorship of "Sex related content for a non-erotic purpose i.e. that is intended to educate"; "Content relating to a social justice movement, for example feminism or anti-racism", "Content relating to minority identity experience i.e. queer content, content about Black experiences", "Self-referential use of slur i.e. d*ke by a lesbian" and "Hate speech" (this latter measurement variable had an opposing effect to the others). We predicted this would cause greater use of obfuscation techniques (H1C). We also included an interaction effect between history of posting controversial content and perceived algorithmic censorship on obfuscation use - those who post controversial content but do not believe this to be censored may be unlikely to use obfuscation.

Perceived effectiveness of techniques was formed of three ratings: perceived effectiveness of "Word substitution (meaning or sound) (sex work → shrek work)";"Spelling changes (LGBT → leg booty)" and "Sound changes with repetition (mother fucker → mugga chugga)" (the remaining six obfuscation techniques identified by Calhoun and Fawcett (2022) did not have significant weights or loadings). We anticipated this would moderate the relationship between perceived algorithmic censorship (of controversial

topics) and use of obfuscation techniques, and between controversial content posting and use of obfuscation technique (i.e. the two motivations related to avoiding censorship through use of obfuscation).

We also anticipated a relationship between posting controversial content and history of reported censorship. We anticipated that history of censorship would then cause greater perceived censorship (of marginalised experiences). History of censorship was formed of two variables: one for the sum of the frequency of removal across topics and another for the sum of the frequency of suppression. We felt summing frequency across topics was appropriate as those who are frequently censored regardless of what they post will likely perceive themselves to face greater censorship than those who are always censored but only for a single topic.

We predicted greater positive emotions associated with obfuscation use (due to the opportunity to express community membership) and stronger social identity community motivations for marginalised respondents, and that these would additionally predict use of obfuscation techniques (H1D). "Positive emotions associated with obfuscation techniques" was reflected in positive emotions ratings (we did not include reversed negative emotions as this reduced construct reliability; we removed ratings for "amused, fun-loving, silly" and "interested, alert, curious" as these had loadings below .708, in line with best practice (Hair et al., 2019)). "Community motivations" was reflected in the ratings given to the five social identity community questions.

To establish the impact of posting controversial content on obfuscation use, we included a latent variable to represent how often the respondent posts controversial content, formed from frequency ratings across the topics. Originally we included all topics, but some were removed in line with PLS-SEM best practice (non-significant weights or loadings), leaving six: "Political content"; "Content insulting or criticising dominant group"; "Content relating to a social justice movement..."; "Content relating to minority identity experience i.e. queer content, content about Black experiences"; "Curse words" and "Content about violence that is not intended to shock or disgust". We anticipate that frequency of posting marginalised controversial topics will be the strongest factor in predicting use of obfuscation techniques, as this has been noted in informal research by journalists, and that the other factors we identify will account for a statistically significant but comparatively small amount of variation.

Our original model had 11 latent variables and 57 observed variables. Given the complexity of this model, we used the "rule of thumb" of 10 respondents per observed variable (for example in Barclay et al. (1995)), giving a minimum sample size of 570

respondents. Whilst such heuristics have been criticised (Westland, 2010; Wolf et al., 2013), we note that this sample size would be considered sufficiently large to detect an effect of "moderate" size per Soper (2025).

7.5.2 Respondents

There were 627 respondents to our survey. The modal age was 23. The majority were female ($60\%, n = 377$), 224 being male (36%), 24 non-binary, 1 "Other"; 1 declined to answer. Twenty respondents were trans (3.0%), eight declined to answer. The majority of respondents were straight ($73.5\%, n = 461$), 44 gay (7.0%), 99 bisexual (15.8%), 8 asexual (1.3%), 4 "Other" (0.6%), 11 declined to answer (1.7%). Seventy-three respondents were disabled (11.7%), eleven declined to answer (1.8%). The majority of respondents were white ($n = 511, 81.5\%$), followed by Black (British) ($n = 45, 7.2\%$) and (British) Asian ($n = 45, 7.2\%$). Twenty-one identified as Mixed/multiple ethnic groups (3.4%), three as "Other" ("Arab", $n = 2$, and "Latine") and two declined to answer.

For the SEM, we removed respondents who did not provide the complete set of demographic information ($n = 25$).

Respondent demographics are also discussed in Chapter 6.

7.5.3 Reliability and Robustness

The reflective measurement models ("positive emotions associated with obfuscation" and "social identity community motivations") all have loadings above .708. Composite reliability was .93 for Social Identity Community, which is a sign of good reliability. For positive emotions, it was .94. Average variance extracted (AVE) was above 0.5, which is a sign of good reliability (Hair et al., 2019). The observed variables in the reflective measurement models are all quasi-metric scales (Hair et al., 2013). Heterotrait-monotrait (HTMT) ratios were all very low (< 0.4) indicating good discriminant validity.

We confirmed that none of the formative variables have a collinearity problem, with variance inflation factors (VIF) all being under 3.0. Loading size and significance was used to determine which observed variables we included in the final model, see above. VIF was satisfactory for all latent variables. Confirmatory tetrad analysis (CTA) supported the finding that the reflective measurement model for positive emotions associated with obfuscation use "[stemmed] from the same domain" (Hair et al.,

2019). Some of the CTA for social identity community ratings *were* significantly different from 0, suggesting this may be more suited to a formative model specification (Hair et al., 2019). However the high VIF and our own theoretical reasoning leads us to consider social identity community more appropriate as a reflective model, in that the ratings reflect a latent sense of belonging.

R^2 was very low ($< .1$) for the endogenous variables of controversial content posting, perceived censorship, positive emotions associated with obfuscation use and social identity community. R^2 was low ($.1 < x < .2$) for experience of censorship. This means marginalised identity accounted for very little of the variation in these variables.

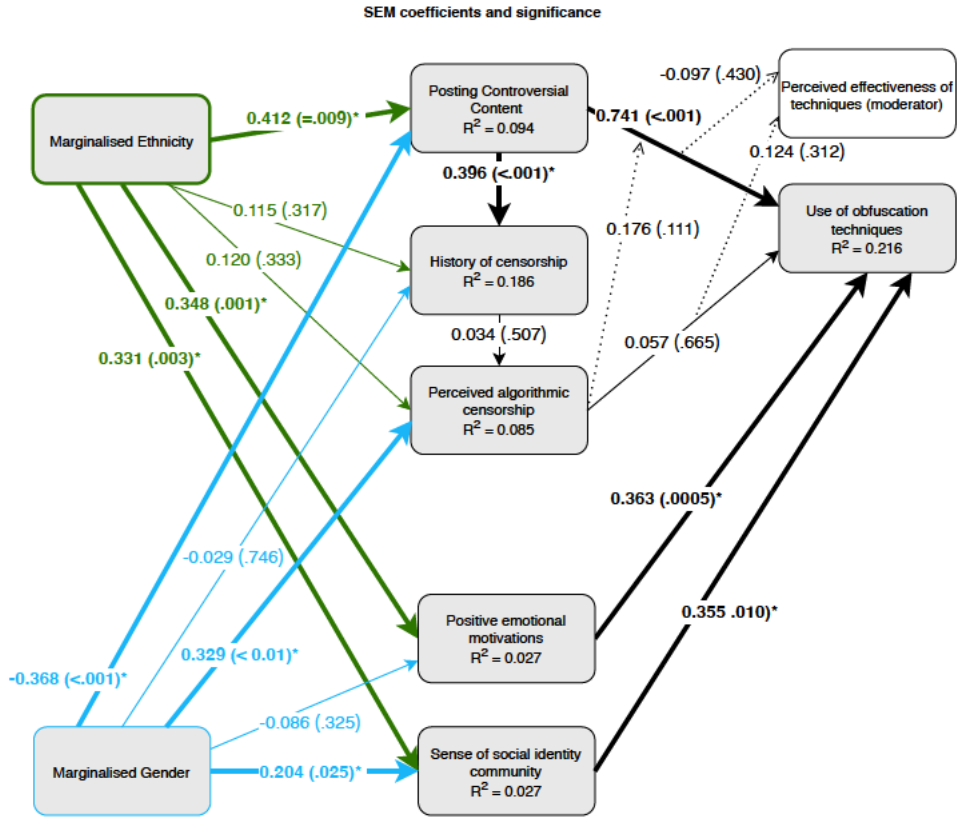
R^2 was also low for our main variable, per our logistic regression (based on latent variable values). Thus whilst we have been able to identify some of the factors influencing obfuscation use (posting controversial content, social community and positive emotions associated with obfuscation use) we have not been able to explain much of the variation. However, low R^2 values are typical in research predicting human behaviour, given the complexity of the phenomena, and given our significant findings, we believe our model captures a substantive difference in behaviour between user groups.

We handled missing data with mean replacement. Missing data was largely ad hoc, due to respondent inattentiveness. 12 respondents reported having never seen obfuscation use and so do not give ratings as to their feelings associated with obfuscation use. We had no evidence that these respondents were substantively different from other users of the app, so we felt mean replacement was appropriate (i.e. if they had seen obfuscation they would approximately share the mean beliefs).

7.5.4 SEM Results

Below we present the results for our SEM and logistic regression analyses for use of obfuscation across gender, sexuality, disability status and ethnic group. The final model including path coefficients and p-values for latent variable relationships, and R^2 value for latent variables is included as Figure 7.3. Path coefficients directly connected to obfuscation use are taken from the logistic regression. Note we only visualise gender and ethnicity in the full model for clarity of presentation. Full results are given in the following in text, and we summarise the results of our hypothesis testing in Table 7.1. In the logistic regression, moderation effects were modelled as interaction effects.

Figure 7.3: Coefficients and R^2 values for final model. Arrowheads indicate a **direct effect**. Dots indicate a moderation effect. For the sake of legibility, we exclude the variables and associated paths for disability and sexuality, which did not demonstrate social identity motivations for obfuscation use. Green is used for effects related to marginalised ethnicity. Light blue for marginalised gender. * indicates significance of $p < .05$. Bold font and increased line thickness signifies significant effects.



Hypothesis		Result
H1A	Minority groups will be more likely to use obfuscation techniques than non-minority groups	Ethnicity: **supported** Gender: not supported Disability: not supported Sexuality: **supported**
H1B	Minority users will (perceive themselves to) face additional censorship (history of censorship) Minority users will (perceive themselves to) face additional censorship (overall censorship)	Ethnicity: **supported** Gender: not supported Disability: **supported** Sexuality: not supported Ethnicity: not supported Gender: **supported** Disability: not supported Sexuality: **supported**
H1C	Higher perceived censorship will predict greater use of obfuscation techniques	Not supported
H1D	Minority users have other motivations for using obfuscation techniques than avoiding automated censorship, related to social identity	Ethnicity: **supported** (sense of community; positive emotions) Gender: **supported** (sense of community) Disability: not supported Sexuality: not supported
H2	Perceived effectiveness of the obfuscation technique moderates the effect of perceived censorship and controversial content posting on the use of the obfuscation technique	Not supported

Table 7.1: Table summarising our hypotheses and whether they are **supported** by the results of our SEM & logistic regression analysis. For hypotheses relating to minority identity.

7.5.4.1 Hypothesis Testing

7.5.4.1.1 H1A: Minority groups will be more likely to use obfuscation techniques than non-minority groups We found that in line with our hypothesis, those marginalised due to their ethnicity (the "not white" group) were significantly more likely to use obfuscation techniques. Per our logistic regression, the effect was significant ($\beta = 0.752, t = 7.254, p = .007$).

In line with our prediction, marginalised sexuality significantly predicted obfuscation use ($\beta = 0.666, t = 5.637, p = .018$)

Counter to our hypothesis, we found that those marginalised due to their gender (the "not men" group) were not significantly more likely to use obfuscation. Examining our model it seems likely this is because whilst those of marginalised genders had stronger social community motivations (which significantly increases obfuscation use) they were also less likely to post controversial content, which would predict less obfuscation use.

Disability status also did not predict obfuscation use nor any of the mediating variables.

7.5.4.1.2 H1B: Minority users will (perceive themselves to) face additional censorship Counter to our hypothesis H1B, we found no relationship between marginalised ethnicity and perceived censorship across topics. We found an effect of ethnicity on history of censorship, ($\beta = 0.278, t = 2.140, p = .032$), which seemed to be mediated by history of posting controversial content ($\beta = 0.163, t = 2.079, p = .038$)

The relationship between gender and history of censorship was significant, such that those marginalised by gender were slightly less likely to have experienced censorship ($\beta = -0.175, t = 2.077, p = .038$). In contrast, we find a significant relationship between marginalised gender and perceived censorship across controversial topics such that the "not men" group were significantly more likely to rate these topics as censored by TikTok. The total effect was significant ($\beta = 0.323, t = 3.460, p = .001$).

Being marginalised by a disability does significantly predict history of censorship ($\beta = 0.410, t = 2.349, p = .019$), but not perceived censorship across topics.

Being marginalised by one's sexuality significantly predicts perceived censorship across topics ($\beta = 0.451, t = 3.992, p < .001$), but not history of censorship.

This suggests that there is some support for our hypothesis in that those marginalised due to their ethnicity or disability perceive themselves to personally face additional censorship. Those marginalised by their gender were actually less likely to report

history of censorship. However, it seems those marginalised due to their gender do perceive there to be greater censorship on the platform in general, likewise do sexual minorities.

7.5.4.1.3 H1C: Higher perceived censorship will predict greater use of obfuscation techniques

Perceived censorship did not predict use of obfuscation, per our logistic regression. Relatedly, there was no moderation effect of perceived censorship on the relationship between posting controversial content and obfuscation use.

7.5.4.1.4 H1D: Minority users have other motivations for using obfuscation techniques than avoiding automated censorship, related to social identity

A primary goal of our SEM was to establish whether factors related to social identity also motivated use of obfuscation techniques by minority users. We found a significant positive relationship between sense of social community on TikTok and obfuscation use ($\beta = 0.355, t = 6.611, p = .010$). Likewise, we found a significant positive relationship between associating positive emotions with obfuscation, and obfuscation use ($\beta = 0.363, t = 7.976, p = .005$).

Looking at our SEM model, we see that gender and ethnic minorities have a significantly greater sense of community on TikTok: $\beta = 0.204, t = 2.247, p = .025$ for marginalised gender; $\beta = 0.348, t = 3.338, p = .001$ for marginalised ethnicity. Those marginalised by their ethnicity also had significantly stronger positive associations with obfuscation use, $\beta = 0.331, t = 3.003, p = .003$.

To summarise, positive emotions and sense of social community predict use of obfuscation, and these social factors are greater for those of marginalised ethnicities. Those marginalised by their gender have a stronger sense of community (which predicts obfuscation use) but not positive emotions. These social factors have no impact on use of obfuscation by those marginalised by their sexuality or disability status.

7.5.4.1.5 H2: Perceived effectiveness of the obfuscation technique moderates the effect of perceived censorship and controversial content posting on the use of the obfuscation technique

We did not find perceived effectiveness to significantly moderate the relationship between perceived censorship and use of obfuscation, nor history of posting controversial content and use of obfuscation.

7.5.4.2 Other SEM Results

We found controversial topic posting predicted history of censorship ($\beta = 0.396, t = 4.641, p < .001$), somewhat trivial given the design of our model (respondents could indicate how frequently other types of content were removed and suppressed beyond the categories we gave, which was included in our measurement variables, but very few did so).

Frequency of posting controversial content significantly predicted frequency of obfuscation use ($\beta = 0.145, t = 6.934, p < .000$), suggesting our pre-selected controversial content types do drive a significant (if not substantial, given the small R^2 value of 0.216) amount of variation in use of obfuscation.

History of censorship did not influence perceived censorship, which suggests that experiencing of censorship do not determine beliefs about wider algorithmic censorship across the platform (we see this starkly in the data related to gender minorities, where those marginalised by gender are less likely to have experienced censorship but more likely to believe in algorithmic censorship across the controversial topics). This reflects our findings in Chapter 6.

7.6 Descriptive Results

In order to further explore R1 (How do motivations for use of obfuscation techniques differ between identities?), we also conduct exploratory analyses. This includes exploring how experiences of obfuscation, such as awareness and comprehension, differ between demographic groups, as these are likely to influence use. We also include analysis of our results for **All Respondents**, as this provides vital context for our comparative analyses. The structure of this section reflects the survey design. We briefly overview experiences of censorship, before focusing on awareness and use of obfuscation use, motivations for obfuscation use and finally social identity. To ensure our results for ***All Respondents*** are indicative of the typical TikTok population, we use raking to weight data by respondent sexuality[8] per their prevalence in the *original* recruitment drive,[9] rather than after up-sampling of LGBTQ+ respondents (as in the previous Chapter). Typically, percentages are given as is appropriate for weighted data,

[8]The ANES raking variable selection algorithm determined changes to the sample based on gender and trans status were negligible

[9]For lack of more detailed TikTok user demographic information being available, we take our original sample distribution to be reasonably accurate

but where it improves clarity we also report raw *n* counts. To support our key claims – those which we return to when discussing the implications of our findings in Section 7.7 – we conduct post-hoc statistical analyses. For clarity, we highlight significant results in bold and summarise them at the end of this Section in 7.6.6.

7.6.1 Respondents

There were 627 respondents. Detailed demographic information can be found in Section 7.5.2. In the following analysis, we do not typically include findings for "Other" sexuality, gender or ethnicity, as these groups are too small and non-homogeneous for meaningful analysis.

The majority (74.1%) of respondents posted on average 0-3 times a month, though all had posted at least once. A recent survey found that for TikTok users who had posted content (around 50% of users), they had posted on average only 6 videos total, suggesting our sample is reflective of typical TikTok users (Bestvater, 2024).

7.6.2 Experience of Censorship

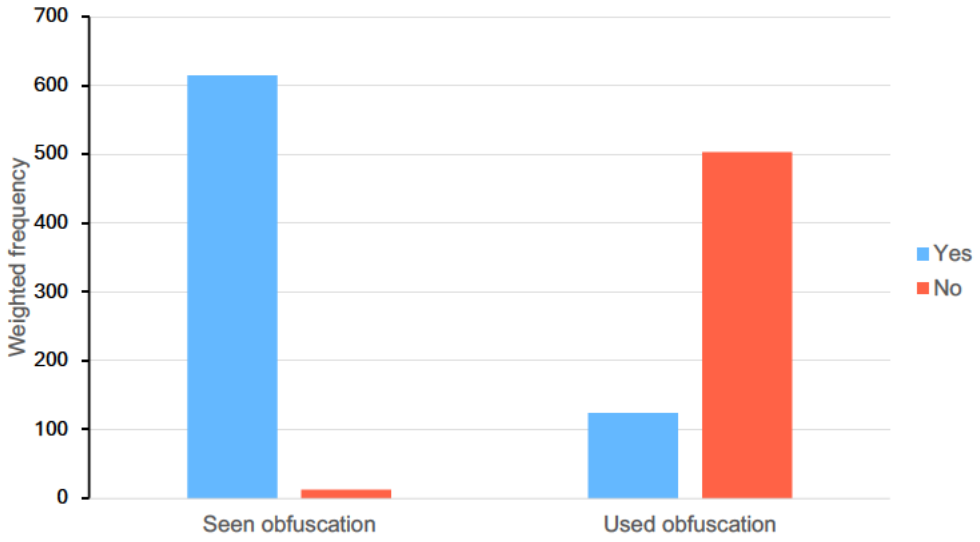
Only around 1 in 8 respondents reported having had content removed (12.8%), and 1 in 7 reported suppression (14.1%). Around half of those who reported having had content suppressed believed they had had content removed, and vice versa. Detailed information on the types of content respondents reported being censored can be found in Chapter 6, along with a breakdown by demographic group.

7.6.3 Awareness and Use of Obfuscation

7.6.3.1 All Respondents

7.6.3.1.1 Familiarity with obfuscation Respondents reported seeing sound changes (82.0%), use of non-letters (8.30%) and word substitution (74.5%) most frequently. Just 2.0% of respondents reported never having seen the obfuscation techniques (and none of these reported seeing any other techniques), and 1.7% were unsure, suggesting that obfuscation is as ubiquitous a phenomenon as media reports and our own experience of TikTok would suggest. 2.0% of respondents reported having seen techniques other than those identified by Calhoun and Fawcett (2022), although upon inspection all would be subsumed under the given categories. When asked whether there was anything else about obfuscation they would like to report, one user said "people on live

Figure 7.4: Bar chart showing the weighted proportion of respondents who have seen obfuscation and who have used obfuscation, demonstrating that whilst awareness is very high, usage is comparatively low.

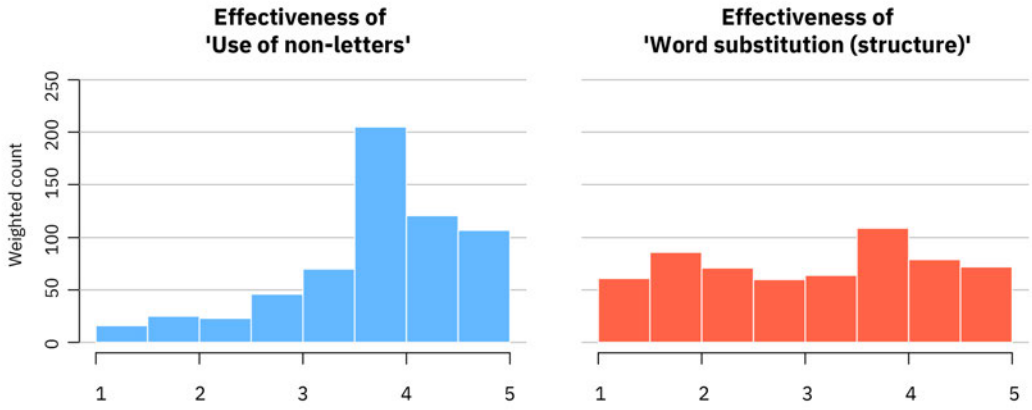


streams... say '61 backwards' instead of 16, to avoid getting banned". In this Chapter, we focus on text-based obfuscation.

7.6.3.1.2 Comprehension of obfuscation Respondents reported understanding the intended meaning "Most of the time" with the greatest frequency (65.6%) followed by "Always" (14.6%). That intended meaning is sometimes not understood suggests the techniques go someway to stopping both "the algorithm" and outgroup members from understanding what is being said.

7.6.3.1.3 Use of obfuscation Despite obfuscation's prevalence on the platform, the majority of respondents reported never having used any of the obfuscation techniques themselves (80.3%). We compare the proportion of respondents who have seen obfuscation vs. have used it themselves in Figure 7.4 . This is perhaps reflective of the majority of our respondents infrequently posting on the platform, as we found a positive (though not linear) relationship between posting frequency and obfuscation use. Only four respondents reported using some other technique, two of which were examples of the existing categories; the novel obfuscation technique was to play music louder than their voice. In our analysis below we take a response of none to mean no obfuscation is used (one respondent selected "Other" and "None" but gave no details so we consider it reasonable to include them as a non-user).

Figure 7.5: Histogram of ratings from “Not effective at all” (1) to “Extremely effective” (5) for “Use of non-letters (gay → 💅)” in blue, and “Word substitution (structure) (homophobia → cornucopia)” in red, showing that respondents generally agreed non-letters were very effective but opinion was divided for word substitutions.



The most popular obfuscation techniques were “Use of non-letters (gay → 💅)” (9.2% of respondents, 46.5% of obfuscation users); “Innovative subword combinations (unalive [un + alive]) (7.6% of respondents, 39.1% of obfuscation users); “Sound changes (sexy → seggsy)” (7.1% of respondents, 38.5% of obfuscation users).

7.6.3.1.4 Effectiveness of obfuscation The obfuscation techniques were all rated between moderately and very effective on average, with “Use of non-letters (gay → 💅)” being rated the most effective on average (3.8). This also had the lowest variance. There was a trend of greater variance being associated with a lower average score, showing opinion was more divided for some techniques e.g. the technique rated as least effective on average was “Word substitution (structure) (homophobia → cornucopia)’ (3.1), which also had the highest variance. We demonstrate different patterns of rating in Figure 7.5. Ratings were slightly higher for all methods across those who use obfuscation - perhaps unsurprisingly, as we would expect those who use obfuscation in order to avoid censorship to believe it is effective at achieving this goal.

7.6.3.1.5 Emotions associated with obfuscation On average, respondents most strongly agreed that using or seeing obfuscation techniques made them feel “interested, alert, curious”, however the mean was only 3.2, just slightly above “Neither agree nor disagree”. Respondents also very weakly agreed with “amused, fun-loving, silly” (3.2). Respondents disagreed with all other emotional statements, and were mostly

likely to disagree with the negative emotions such as shame, fear, disdain or guilt. But even these were not strong disagreement, and it seems obfuscation techniques do not typically elicit strong emotions. This was equally true when considering only those who actively use obfuscation.

7.6.3.2 Exploratory

7.6.3.2.1 Familiarity with obfuscation Awareness of different obfuscation techniques was similar across genders, disability status and ethnicity. Trans respondents were much more likely to have seen spelling changes, perhaps reflective of the example given ("LGBT $\rightarrow$ leg booty") which is more relevant to the trans experience, a pattern also evident for bisexual, gay and asexual respondents compared to straight respondents.

7.6.3.2.2 Comprehension of obfuscation Considering comprehension, we saw similar ratings for comprehension across male and female respondents, with the small number of nonbinary respondents giving the highest comprehension ratings on average: nonbinary respondents all understood the intended meaning at least half the time, whereas around 10% of men and women only understood "Sometimes" or "Never". We noticed a similar pattern for trans versus non-trans respondents, and for LGB+ versus straight respondents. **A Pearson's χ^2-test comparing members of the LGBTQ+ community to those outwith the community found a significant difference in levels of comprehension**, $\chi^2(5) = 13.75, p = 0.017$. This suggests obfuscation may at times be acting as a form of modern polari (Baker, 2019), an anti-language that is deliberately illegible to outsiders as well as algorithms.

Respondents with disabilities were slightly more likely to understand the intended meaning than those without, and we saw a similar pattern for non-white compared to white respondents, with some variation across ethnic groups.

7.6.3.2.3 Use of obfuscation Looking at use of obfuscation techniques, we found the nonbinary respondents were more likely to report having used obfuscation techniques (41.7% of nonbinary respondents, $n = 10$) compared to men (24.1%, $n = 54$) and women (16.2%, $n = 61$). **A Fisher's Exact Test (appropriate to the small number of non-binary respondents) found non-binary respondents were more likely to use obfuscation than those of binary genders,** $p = .015$. We return to the impact of gender on obfuscation use in Section 7.5.

Ethnic Group	Used obfuscation
(British) Asian (45)	28.9% (13)
Black (British) (45)	42.2% (19)
Mixed/ multiple ethnic groups (21)	38.1% (8)
White (511)	16.4% (84)

Table 7.2: Table showing the use of obfuscation techniques by ethnic group. Percentages represent the number of respondents who did *not* select "None" from the list of obfuscation techniques they had used. Counts given in brackets.

The particular obfuscation techniques used differed across genders e.g. nonbinary respondents were much more likely to report using word substitution based on meaning or sound (16.6%, $n = 4$) compared to men (8.0%, $n = 18$) or women (4.0%, $n = 15$), and also much more likely to report use of non-letters (29.2%, $n = 7$) compared to men (8.5%, $n = 19$) or women (9.0%, $n = 34$). This suggests different norms may exist within binary and nonbinary gender identity communities.

Similarly, trans respondents reported using obfuscation at a higher rate (30%, $n = 6$) compared to non-trans respondents (19.5%, $n = 119$), though this difference was not significant per a Fisher's Exact test. The very small number of trans respondents who use obfuscation make it hard to identify more fine-grained trends.

Looking across different sexuality demographics, we noticed some differences in the use of different obfuscation techniques. Straight respondents were less likely to report using obfuscation (16.3% of straight respondents ($n = 75$), compared to gay (22.7%, $n = 10$), bisexual (35.4%, $n = 35$) and asexual (25%, $n = 2$) respondents). **A Pearson's χ^2-test comparing members of the LGB+ community to straight respondents found a significant difference in use of obfuscation,** $\chi^2(1) = 14.77, p < .001$. Considering particular types of obfuscation, LGB+ respondents were more likely to report use of non-letters compared to straight respondents.

Differences across gender identities, trans history and sexuality suggest that LGBTQ+ respondents use different obfuscation techniques compared to cisgender and/or straight respondents, and at a higher rate.

Respondents with disabilities were more likely to use obfuscation techniques (27.4%, $n = 20$) compared to those without disabilities (19.0%, $n = 103$), though this difference was not significant per a Chi-square test. There was a marked difference in use of some techniques between groups, for example 16.4% ($n = 12$) of respondents with

disabilities reported using non-letters compared to 8.3% ($n = 45$) of those without.

Respondents of colour were more likely to use obfuscation techniques than white respondents, as shown in Table 7.2 (we return to the impact of ethnicity on obfuscation use in Section 7.5). We noticed a number of differences in the use of specific techniques across ethnic groups, for example Black (British) respondents were the most likely to use sound changes (20.0%, $n = 9$), almost twice as likely as any other group. This suggests respondents belonging to different ethnic groups may use distinct obfuscation techniques within their (online) cultures.

7.6.3.2.4 Effectiveness of obfuscation We noticed some differences in the perceived effectiveness of different obfuscation techniques across gender and sexuality groups, perhaps reflective of their direct experiences of censorship on the platform. For example, nonbinary respondents felt that use of non-letters was more effective compared to men and women. Gay respondents typically rated all techniques as less effective compared to straight respondents (barring use of non-letters which gay respondents rated as slightly more effective), despite gay respondents being more likely to use obfuscation compared to straight respondents. It seems perceived effectiveness is not a strong predictor of obfuscation use - perhaps because obfuscation plays more roles than simply "tricking the algorithm".

7.6.3.2.5 Emotions associated with obfuscation Looking at those marginalised by their gender, we noted that differences between the groups were most obvious when we focused on those who report using obfuscation themselves. The following results are for those who use obfuscation. Men agree more strongly than women and nonbinary respondents that obfuscation makes them feel amused (3.6 compared to 3.3 and 3.1 respectively) (the only statement to which all three gender groups did not disagree). Men were much more likely to agree to the positive emotions associated with obfuscation use, such as hopeful, inspired, interested and proud. Women agreed to also feeling inspired, and nonbinary respondents disagreed to all other positive emotions. Thus amongst those who use obfuscation, those marginalised by their gender are much less likely to feel positive emotions associated with obfuscation.

Trans respondents typically disagreed less strongly with negative emotions and agreed less strongly with positive emotions, suggesting overall they have more negative associations with obfuscation.

We found that bisexual respondents responded similarly to the emotions associ-

ated with seeing obfuscation use when compared to straight respondents, whilst gay respondents showed some differences, for example gay respondents disagreed to feeling amused by obfuscation (2.9), in contrast to straight and bisexual respondents who weakly agreed, bisexual respondents most strongly (3.3 compared to 3.2 for straight). Thus, as with patterns of use of obfuscation, we notice differences across marginalised sexualities.

Looking across all respondents by disability status, we found differences in agreement with positive emotions were insubstantial. When we focused on those who use obfuscation, we found those with disabilities were less likely to agree to positive emotions.

We found that compared to other ethnic groups, black respondents associated more positive emotions with obfuscation, and this was particularly evident when looking at those who actively use obfuscation: this is illustrated in Table 7.3, which shows that Black (British) respondents who use obfuscation agreed more strongly with the positive emotions on average, compared to other ethnic groups. We return to the topic of ethnicity, emotions and obfuscation use in Section 7.5.

Ethnicity	amused	awe	content	glad	grateful	hopeful	inspired	interested	love	proud
(British) Asian	3.5	3.2	2.8	3.3	2.8	3.0	3.0	3.3	3.0	3.1
Black (British)	**3.8**	**3.3**	**3.6**	**3.6**	**3.5**	**3.7**	**3.5**	**3.8**	**3.4**	**3.3**
Mixed/multiple	3.1	2.7	3.1	2.9	3.0	2.8	2.3	3.4	2.2	2.5
White	3.4	2.9	2.6	3.0	3.0	3.0	2.8	3.4	2.7	2.7

Table 7.3: Table showing average agreement rating with each of ten positive emotions. **Bold** indicates the strongest agreement.

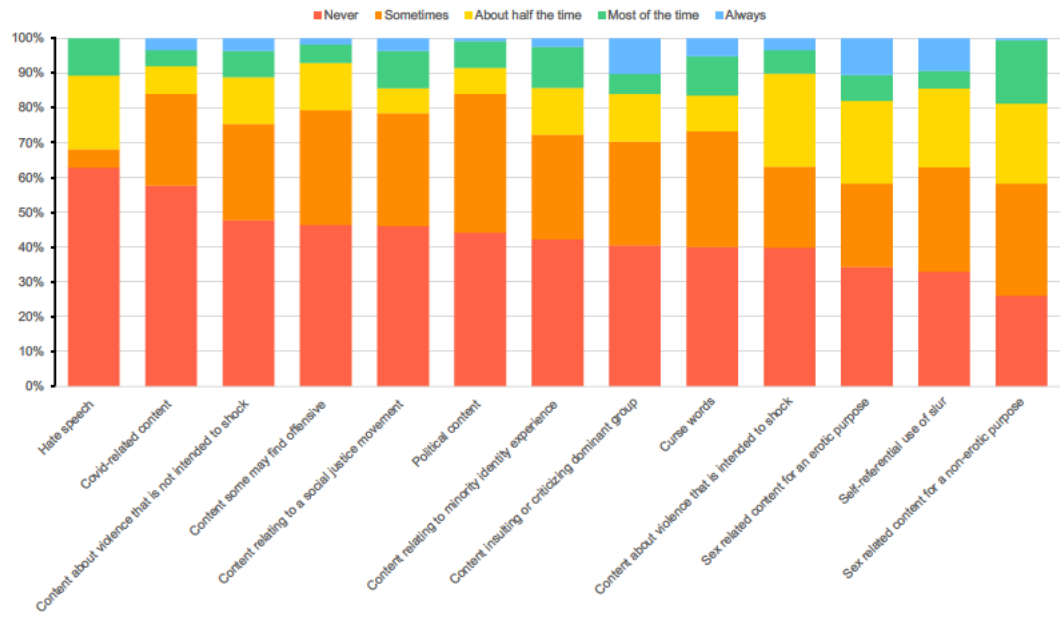
7.6.4 Motivations for Obfuscation Use

7.6.4.1 All Respondents

The least common of the controversial content types that respondents posted was “Hate speech” (3.2%). The most common controversial content types were “Curse words”, with 204 respondents reporting posting this (35.1%). A detailed analysis of the types of content posted can be found in Chapter 6.

We visualise use of obfuscation across content types in Figure 7.6. For “Hate speech” and “Covid-related content”, the majority ($> 50\%$) of respondents who posted this kind of content reported “Never” using obfuscation techniques. For all other

Figure 7.6: Bar chart showing the weighted proportion of respondents who used obfuscation at each frequency (from "Never" to "Always") across content types, ordered by frequency of "Never" using obfuscation, demonstrating how obfuscation use differs substantively across content types.



content types, the majority of respondents reported using obfuscation at least "Sometimes". This suggests that obfuscation is widely used across the types of content typically subject to social media censorship, though to differing degrees. This was true even for topics which did not go against community guidelines, such as content relating to minority identity experiences. The content type where obfuscation was most likely to be used was "Sex related content for a non-erotic purpose i.e. that is intended to educate", with 41.7% of respondents reporting using obfuscation at least half the time. This is reflective of findings of excessive moderation of sex education content reported elsewhere (Are, 2023).

Very few respondents reported using obfuscation techniques for any topics other than those we had pre-identified, specifically related to mental health issues ($n = 2$, raw count), and abstract or gaming mentions of death ($n = 3$). Thus our pre-selected categories, derived from the social media censorship literature, appear to be a good match for the types of content associated with obfuscation use. Two respondents mentioned using obfuscation on other social media sites.

Looking at those who posted obfuscation techniques, we found that respondents somewhat agreed with the statement "So the algorithm does not censor my posts"

(4.0), and very weakly agreed with "Because it's fun" (3.2) and "To protest algorithmic moderation" (3.2), and for all other statements there was weak disagreement on average ($3 > X > 2$), including those statements that related to use of coded language to signal and build community. However for all motivations there was considerable variance.

We invited respondents to write anything else they would like on their use of obfuscation techniques. Some interesting comments included "for me obfuscation techniques have almost moved away from censorship into slang in some communities" and "I use obfuscation in those cases to avoid censorship and to keep my tone a little more upbeat than in my academic research to match the dominant tone on TikTok (which sometimes does feel a bit syrupy sweet and artificial)." These respondents demonstrate some meta-awareness of their motivations for obfuscation use. Another respondent answered "I only use it to avoid being punished by the tiktok algorithm but I hate it", suggesting that whilst some communities associate positive emotions with obfuscation use (see above) this is far from universal.

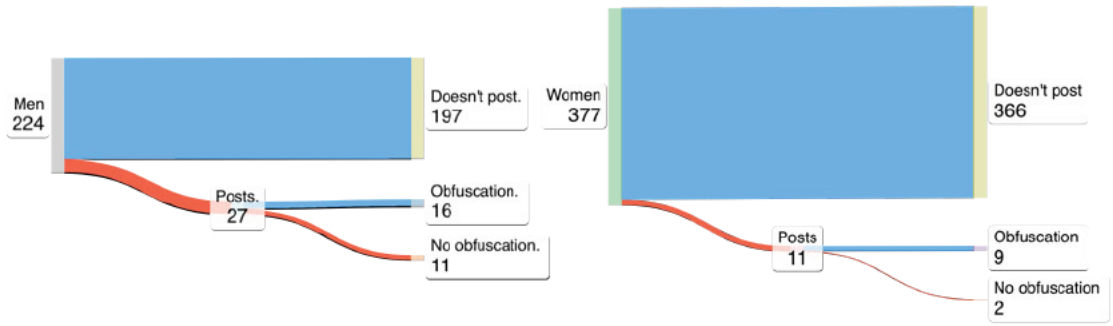
7.6.4.2 Exploratory

An analysis of how the types of content posted differ across demographic groups can be found in Chapter 6.

Looking at differing use of obfuscation techniques between demographic groups, we turn first to differences by gender. One striking difference was that when women post erotic content they are much more likely to use obfuscation technique (81.8% of women who post erotic content, $n = 9$) compared with men (59.2%, $n = 16$), as shown in Figure 7.7. **A Pearson's χ^2-test comparing men and women found a significant difference in the frequency of use of obfuscation techniques,** $\chi^2(5) = 25.56, p < 0.001$. This could reflect differences in the (perceived) "policing" of women's sexuality compared to men - a well documented phenomenon on TikTok and other platforms (Are, 2023, 2020). The low number of nonbinary respondents posting controversial content makes it difficult to identify trends. Similarly we were unable to identify trends for trans respondents.

Turning to sexuality, we noticed some differences in the use of obfuscation techniques. For example, bisexual respondents were much more likely (75%, $n = 36$) to use obfuscation techniques when posting curse words compared to gay (50%, $n = 7$) and straight (55.5%, $n = 83$) respondents, perhaps reflecting different trends in the community and/or a greater fear of censorship, despite curse words not being explicitly

Figure 7.7: Sankey chart showing the count of respondents who post erotic content and use obfuscation versus never use obfuscation when doing so, by gender, demonstrating that men are more likely to post erotic content and less likely to use obfuscation.



(a) Posting of erotic content by men (b) Posting of erotic content by women

Ethnic Group	Minority experience	Social Justice	Crit. dom. group
(British) Asian	46.1 (6)	42.9% (6)	66% (4)
Black (British)	92.9% (13)	81.3% (13)	90% (9)
Mixed/ multiple ethnic groups	50% (2)	60% (3)	100% (2)
White	48.3% (28)	48.7% (38)	43.8% (14)

Table 7.4: Table showing the use of obfuscation techniques for specific topics by ethnic group. "Crit. dom. group" refers to content criticising a dominant group. Percentages represent the number of respondents who reported at least "Sometimes" using obfuscation when posting this topic. Number of respondents given in brackets.

prohibited by community guidelines.

Black respondents were more likely than white respondents to use obfuscation techniques for all content types other than curse words where usage was on par, reflecting their greater use of obfuscation overall. We noticed substantial differences in the reported use of obfuscation for content relating to a minority identity experience, social justice movements, or which criticises a dominant group, across ethnic groups, as shown in Table 7.4: Black respondents stand out as being much more likely to report consistently using obfuscation techniques when posting these types of content, when compared to other ethnic groups.

Considering differences between respondents with and without disabilities, we find that generally behaviour is similar, with one noticeable difference being that those

without disabilities are more likely to use obfuscation when posting content some may find offensive (56.4%, $n = 53$) compared to those with disabilities (35%, $n = 7$). This might reflect differences in the type of potentially offensive content that these two groups are posting, or else it may reflect different beliefs about censorship (we discuss some differences in Section 6.4.2.5).

Looking at motivations for obfuscation use (by those who reported ever using obfuscation) we found that women and nonbinary respondents agreed to using obfuscation to protest algorithmic moderation (3.3 and 3.6 respectively) whilst men neither agreed nor disagreed (3.0). Women and nonbinary respondents also agreed more strongly that they use obfuscation to avoid censorship (4.1 and 4.2 respectively) compared to men (3.8). Men agreed they used obfuscation "because it's fun" (3.3), women neither agreed nor disagreed (3.1), whilst nonbinary respondents disagreed (2.7). This suggests that the role obfuscation plays differs between genders, although with only ten nonbinary respondents using obfuscation, we hesitate to draw any strong conclusions about differing behaviours between binary and nonbinary gender users of TikTok. Trans respondents disagreed to using obfuscation because its fun (2.7) whilst non-trans respondents agreed (3.2), although again with the small number of trans respondents who use obfuscation we must be cautious in drawing conclusions from this data.

Comparing across sexualities, we notice differences in reported motivations for using obfuscation. Gay and bisexual respondents disagreed to using obfuscation to express creativity (2.1 and 2.7, respectively), whilst straight respondents neither agreed or disagreed (3.0). Straight respondents agreed to using obfuscation "because it's fun" (3.3), where gay and bisexual respondents did not (3.0 and 3.0 respectively). This reflects a similar finding for nonbinary versus binary respondents (and trans versus non-trans, though see above note), suggesting that LGBTQ+ respondents may be less likely to agree that obfuscation use is fun – although we reiterate that we typically had very low counts for queer respondents who use obfuscation.

Nonbinary, trans, bisexual and gay respondents all disagreed to using obfuscation so that only certain TikTok users could understand them (1.9, 2.4, 2.4 and 2.6, respectively). This contrasts with the finding that non-LGBTQ+ respondents were more likely to report not understanding some use of obfuscation. This suggests that whilst most queer respondents may not deliberately be using obfuscation to exclude outgroup members (as they might have done with polari), nonetheless it can play this role unintentionally.

Comparing respondents with and without disabilities, we found they rated most of

the statements very similarly, except for "To show which communities I belong to", where respondents without disabilities disagreed to these statements (2.6) and those with disabilities neither agreed nor disagreed (3.0).

Looking at differences across ethnic groups, we found that Black (British) and (British) Asian respondents weakly agreed that they used obfuscation to feel part of the TikTok community (3.3 and 3.2), whereas those of mixed/multiple ethnic groups and white respondents disagreed (2.1) and white respondents neither agreed nor disagreed (2.9). Black (British) respondents agreed that they used obfuscation "because it's fun" (3.80), which was not true for (British) Asian and (3.1) and white respondents (3.1), and only weakly true for those of mixed/multiple ethnic groups (3.2). These differences suggest that the role of obfuscation, at least as respondents are aware of it, differs across online cultures.

7.6.5 Social Identity

7.6.5.1 All Respondents

We found there was agreement for all the social identity questions on average, with respondents most strongly agreeing to "I feel accepted by members of my community [on TikTok]" (3.6) and least strongly agreeing to "I feel a connection with other members of my communities" (3.4). Amongst those who use obfuscation, agreement rates were always higher, for example these respondents agreed more strongly that "I feel accepted by members of my community [on TikTok]" (3.9).

7.6.5.2 Exploratory

Respondents of all genders agreed to all statements. Women agreed more strongly compared with men. Nonbinary respondents agreed less strongly than men and women that they feel a connection (3.2 compared to 3.3 and 3.5, respectively), feel accepted (3.3 compared to 3.5 and 3.6, respectively) and feel valued (3.3 compared to 3.4 and 3.5, respectively) by members of their community on TikTok, otherwise ratings were similar to those of women.

Asexual respondents agreed most strongly to the social community statements on average, having the highest agreement rating for all statements bar "I feel a connection with other members of my communities" (3.5), perhaps suggesting TikTok has helped many asexual people find community. Gay respondents typical rated statements on par

or lower than straight respondents, whilst bisexual respondents typically rated statements on par or higher than straight respondents.

Trans respondents showed a different response pattern to non-trans respondents. Specifically, they agreed more strongly to all statements, and this was particularly true for "I feel like I am part of my communities" (4.1 compared to 3.5).

Respondents with disabilities less strongly agreed to feeling accepted (3.4 vs 3.6) and valued (3.2 vs 3.5) by their communities compared to those without disabilities - other ratings were similar (± 0.1)

Respondents of colour always agreed more strongly to the social identity statements compared to white respondents.

7.6.6 Summary of Significant Results

In Table 7.5, we summarise the key findings from our descriptive analysis that we determined to be significant, per post-hoc tests of independence. We discuss the implications of these findings contextualised with those of our SEM analysis in Section 7.7.

Finding	Post-hoc test	Result
LGBTQ+ respondents are significantly more likely to understand the intended meaning of obfuscated content compared to non-LGBTQ+ respondents	Pearson's χ^2-test	$\chi^2(5) = 13.75$, $p = 0.017$
Non-binary respondents were significantly more likely to use obfuscation that those of binary genders	Fisher's Exact Test	$p = .015$
LGB+ respondents were significantly more likely to use obfuscation than non-LGB+ respondents	Pearson's χ^2-test	$\chi^2(1) = 14.77$, $p < 0.001$
Women are significantly more likely to use obfuscation when posting erotic content compared with men	Pearson's χ^2-test	$\chi^2(5) = 25.56$, $p < 0.001$

Table 7.5: Summary of significant post-hoc tests of independence when comparing experiences with obfuscation between marginalised and non-marginalised respondents.

7.7 Discussion

7.7.1 Implications for Theory

We set out to address our primary research question: how do motivations for use of obfuscation techniques differ between identities? Our SEM results show a significant relationship between marginalised identity and rate of obfuscation use, one primarily driven by differences in the types of content posted. We also found that social identity community strength and positive emotions impact obfuscation use, and in particular mediate the relationship between marginalised ethnicity and rate of obfuscation use. Counter to our predictions we did not find strength of belief in censorship across controversial topics predicted obfuscation use. We complemented our statistical modelling with descriptive analysis, and below we discuss how together our analyses paint a picture of obfuscation use, particularly by marginalised users of TikTok. In line with media reports (and our experiences as authors), our results show that obfuscation use is ubiquitous on TikTok, with over 95% of our respondents having seen obfuscation on the platform. Use of obfuscation by our respondents was much lower in comparison - only around fifth had ever used obfuscation. The media reports of widespread obfuscation use may be because a small number of individuals, such as popular TikTok creators, use obfuscation techniques heavily. In contrast to these creators, our sample, which represents more typical users of the platform who post infrequently, rarely use these techniques. Use of obfuscation was largely determined by the types of content posted: we found that the controversial topics we had devised based on existing literature seemed to account for almost all obfuscation use, with very few users reporting using obfuscation for any other topics. This reflects the expected relationship between experience of and beliefs about wider social media censorship, and use of obfuscation on TikTok.

We typically found obfuscation use was highest amongst marginalised respondents, except in the case of gender. In our logistic regression analysis we found only sexuality and ethnicity significantly predict obfuscation use. People of colour are significantly more likely compared to white respondents to use obfuscation techniques. LGB+ respondents are significantly more likely to use obfuscation than straight respondents.

In contrast, and counter to our hypothesis, we did not find those marginalised due to their gender to be significantly more likely to use obfuscation techniques compared to men, per our logistic regression. However, considering our descriptive findings, it may be more accurate to say that *women* are not more likely to use obfuscation com-

pared to men, because we actually found obfuscation use to be very high for nonbinary respondents - however they account for only 3% of the sample, so had limited impact on our PLS-SEM model and logistic regression. This also highlights the importance of pairing our SEM model with exploratory descriptive analyses. High use of obfuscation techniques was typical of our LGBT+ respondents. Thus, marginalised identity is typically associated with greater obfuscation use, except in the case of (binary) gender.

History of posting marginalised controversial content was the greatest predictor of obfuscation use, and this was significantly impacted by both ethnicity and gender. People of colour post more controversial content, and thus use obfuscation more, whilst those of marginalised genders (perhaps more accurately, women) post less controversial content (but are no less likely to use obfuscation, perhaps because of the opposing effect of social community strength). As with Haimson et al. (2021), we found that much of the "controversial" content posted by marginalised users often does not go against community guidelines, but these users nonetheless feel the need to use obfuscation to avoid censorship.

Whilst overall perceived algorithmic censorship does not significantly predict obfuscation use per our logistic regression, it is informative to consider the relationship between the perceived censorship of specific topics and obfuscation use. Perceived censorship will likely be informed by the users' experiences of censorship more widely, which may in turn impact their behaviour. For example, when posting types of content which have previously been more heavily policed (literally and metaphorically) for women compared to men, women may be more likely to use obfuscation – as we found to be the case for erotic content, where obfuscation use was particularly high for women compared to men.

A primary motivation for our investigation was to establish whether social motivations (related to sense of community and creative use of language) also influenced obfuscation use by marginalised users, in addition to the types of content being posted. We found that positive emotions associated with obfuscation use did predict greater use of obfuscation techniques, and this mediated the relationship between marginalised ethnicity and obfuscation use. Looking at specific ratings across all respondents, we typically found that "interested" and "amused" were the only emotions that respondents agreed to. Interestingly, these were the positive emotions we removed from our model, in line with SEM best practice, perhaps because they sit at odds with the general trend of ambivalence or disagreement with all emotion terms. From our descriptive data, we typically found that users of obfuscation who were marginalised by their iden-

tity associated less positive emotions with obfuscation, except for Black (British) respondents (see our discussion of playfulness below), and indeed in our model there was a significant relationship between marginalised ethnicity, associating positive emotions with obfuscation, and greater obfuscation use. We found that amongst those who use obfuscation techniques, men were much more likely than women and nonbinary people to associate positive emotions with obfuscation. Those marginalised by their gender may see obfuscation use as more of a necessity, to avoid "unfair" censorship, whilst men see obfuscation use as an opportunity for word play (although women also agreed on average to using obfuscation for fun, despite being less likely to associate positive emotions with obfuscation).

Indeed, we found that across demographic groups, respondents differed as to whether they reported using obfuscation for fun. Whilst we are not the first to suggest obfuscation use is a form of language play (Calhoun and Fawcett, 2022, 2023), our findings highlight how engagement in this kind of play differs across demographics groups. LGBT+ respondents typically did not agree to this statement, in contrast to cis-straight respondents, nor did marginalised sexuality predict positive associations with obfuscation, suggesting that for queer users of TikTok, use of obfuscation is seen as more of a necessity than an opportunity for play. In contrast, Black (British) respondents agreed to using obfuscation for fun, unlike other ethnic groups. Thus in the Black (British) online community, obfuscation may be interpreted as a form of wordplay. This is reminiscent of the way other forms of wordplay play a prominent role in "Black twitter" (Florini, 2014) (a term typically associated with African American twitter users, but which has also been used to refer to a community of Black British users (Knight, 2021; Klassen and Fiesler, 2022)). Use of certain obfuscation styles or techniques may even come to index Black identity, and then be subject to the re-appropriation which is so widespread on TikTok, as Jones (2023) discusses in particular with regards to use of Black voices by white creators.

In our model, social identity community strength had a positive relationship with obfuscation use, supporting our prediction that other factors than avoiding censorship would influence obfuscation use. We found that marginalised respondents (for ethnicity and gender) reported a significantly stronger sense of social identity. Looking at the descriptive data, we note that agreement differed across identity groups, suggesting that whilst TikTok may be used as a platform to express and construct one's identity, as has been found elsewhere, the experience is not always affirming (Simpson and Semaan, 2021; Zeng and Kaye, 2022).

Beyond affirming social identity, obfuscation can act to exclude outgroup members. When asked directly, LGBT+ respondents did not agree that they used obfuscation so that other TikTok users could not understand what they were saying. However, we found that non-LGBT+ respondents were more likely to report not always understanding the intended meaning, suggesting obfuscation use by LGBT+ users may inadvertently act as a modern day Polari (Baker, 2019). We typically found non-marginalised groups were less likely to report always understanding the intended meaning. Whilst excluding outsiders seems not to be a primary motivation for obfuscation use, it nonetheless has this effect. For some marginalised groups – for example Black British users – obfuscation may also be a way to engage in playful language use whilst maintaining an element of secrecy, as with other anti-languages (Halliday, 1976).

It remains to be seen to what extent "platform vernaculars" (Gibbs et al., 2014), such as obfuscation use, affect "offline" language use. Anecdotally, we have heard "nip nops" and "unalive" verbalised in real life, well outside of the reach of any censorship algorithm. Especially where the use of the "platform vernacular" is motivated by language play or to express affiliation with a social group not confined to a digital platform, broader popularisation of expressions originating as obfuscation is very possible.

7.7.2 Implications for Social Media Policy and Industry

We found disparities in perceived censorship across genders, more specifically greater perceived censorship of content related to identity and sex education according to those of marginalised genders, which aligns with previous findings that women face additional scrutiny online (Are, 2020, 2022), and is no doubt informed by women's experiences of censorship elsewhere (as discussed above). Whilst moderation may be necessary (and in the UK, required by law (DSIT, 2025)) to prevent the dissemination of sexually explicit, violent or exploitative content, which will predominantly harm women, the over-moderation of women's content can likewise contribute to the power structures that ultimately enable violence against women. This suggests a role for advocacy groups to monitor, and regulators to enforce, equality standards for the *experienced* treatment of marginalised genders by social media algorithms, just as gender equality is monitored and enforced in another form of media, namely advertising.[10]

[10] https://www.asa.org.uk/advice-online/harm-and-offence-gender-stereotypes.html

Of course, the topic of sexual content moderation invites discussion of how to handle different user perspectives. Whilst preventing extreme sexual content on TikTok may be necessary to comply with local laws, such as the Online Safety Act (DSIT, 2025), users will also have different preferences for how legal sexual content should be handled. Some may wish to prevent any sexually suggestive content on the platform, to align with their religious or moral values, whilst others may favour the value of sexual self expression. Such diverging perspectives will be true for many sensitive topics, such as discussions of suicide, sexual violence, depictions of violence, "identity politics" and countless others. To handle these differing perspectives, researchers have explored personalised moderation interventions (Cresci et al., 2022), including how these might be received by the public (Jhaver et al., 2023). The ubiquitous nature of obfuscation on TikTok suggests dissatisfaction with the current "one size fits all" approach to moderation.

Considering the implications of our findings for the social media industry, our work makes it clear that regardless of the "true" nature of their algorithms, TikTok and other social media companies should be aware of the work marginalised users do – namely, using obfuscation – to feel they have a chance of being treated fairly. Clearly, the impact of "the algorithm" is not uniformly experienced. Any attempt to "de-bias" the moderation and recommender algorithms (TikTok has a history of publicly promising to fix automated discrimination (Brown, 2021; Ohlheiser, 2021)) must also aim to improve the *perceived* fairness of the algorithms, and account for the behaviours that users – particularly marginalised users – are already employing to try and level the playing field. For example, allowing third party auditors to explore how sensitive topics including identity are handled by the moderation and recommender algorithms, and publicly addressing any issues raised, would help to improve perceived fairness more effectively than internal audits even if the outcomes (with regards to modification of the algorithms) are identical. For TikTok in particular, it is important to address the reputational damage that has been done by previous reports of unfair moderation (Lorenz, 2022; Brown, 2021; Bacchi, 2020; Ohlheiser, 2021; Hern, 2019a,b). Current lack of trust is reflected in the high use of obfuscation techniques even for topics which TikTok claims not to moderate, such as posts relating to minority identity experiences.

Failure to account for perceived fairness will mean that user behaviour will continue to undermine the effectiveness of moderation systems. Whilst content creators may be primarily motivated by a desire to redress perceived bias, and in some cases by a desire to express linguistic identity, one negative consequence of obfuscation use is

that other users will be unwittingly exposed to distressing content. For example, a user who has set up keyword filters to avoid being exposed to content relating to "suicide" may nonetheless be served content featuring mentions of "sui" or "sewercide",[11] even though the original creator's intention was only to avoid automated rather than user-driven moderation. Obfuscation use to avoid algorithmic censorship thus also makes it harder for the users themselves, and not just the platform, to control what they are exposed to. If users felt they were being allowed to express themselves authentically, without fear of unsubstantiated censorship, they would be less likely to use obfuscation (cf. the role that obfuscation plays in establishing identity for some communities). This would mean that content that could genuinely harm others (for example, graphic descriptions of suicide) can be more effectively identified (by the platform and by other users). Thus perceived bias in censorship algorithms has implications for the safety of the platform. TikTok and other social media companies must work with users to establish when community guidelines are felt to be unfairly enforced, so users do not resort to their own bias mitigation techniques (such as obfuscation), which make community guidelines harder to enforce, and may lead to the platform casting a wider net with a lower threshold for harm which in turn drives up avoidance behaviours.

Another unintended negative consequence of obfuscation is that it may be harder for users to find vital content or community. For example the grape emoji, 🍇, or a purple dot, 🟣, is often used on TikTok to refer to rape (Croquet, 2024). If users are not familiar with this obfuscation, they will not only find it harder to avoid triggering content, but also be unable to find support from educators or from other users who have experienced rape who are using this obfuscation to avoid censorship. Finally, significant concern has been expressed about the potential for such obfuscations to reinforce the taboos around sensitive subjects. For example, one journalist argued that by fuelling users' suspicions about moderation of content related to sexual violence, such that they feel motivated to use obfuscations, platforms like TikTok are responsible for reinforcing the taboo around discussing this topic (Croquet, 2024).

[11] We found that searching for the common obfuscations "self unalive" and "sewer side" leads to being served a supportive message and links to mental health support, suggesting that the TikTok user experience team are trying to keep on top of common obfuscations, even if it remains unclear how the algorithm handles them.

7.8 Limitations and Future Directions

When modelling posting controversial content, we did not factor in TikTok-specific issues such as copyright, strict minor safety etc. which may explain why so much variation in use of obfuscation is unaccounted for.

We may have gathered more data relevant to obfuscation technique use by recruiting users on TikTok who post content, via promotional material on the app. However, as we note in Chapter 6, this would have introduced an unknown sampling bias into our recruitment. We did not recruit solely those who post "controversial" content, as we wished to understand factors influencing obfuscation use beyond the type of content. Likewise we did not recruit solely those who post frequently, as it was *a-priori* unclear whether obfuscation use motivations might differ across user frequency groups. However, we acknowledge that this may have resulted in us gathering more relevant data.

The specific examples we gave for each obfuscation technique may have heavily influenced whether respondents agreed to having seen or used the general technique. This could explain why use of non-letters was higher for LGBT+ respondents, because the specific example was about obfuscating the word gay. We provided a single example to avoid overwhelming respondents but this may have unduly influenced our findings.

There are user groups who have reported unfair censorship on other social media platforms but who we did not consider in our study, for example users who experience weight discrimination due to being (self-described as) fat (Clark et al., 2021) or users with a history of sex work (Are, 2022). Respondents were given the opportunity to write any other marginalised groups they belonged to, and whilst several identities were submitted we were not able to establish any additional sub-populations to consider e.g. only one respondent wrote in "ex-sex worker".

We are unable to make our data publicly available, as we wanted to minimise the risks of de-anonymisation, given respondents were sharing particularly sensitive information related to their identity, and their use of TikTok to post controversial or "sensitive" topics. However, we acknowledge that lack of data transparency is a limitation of this work, as is the lack of pre-registration.

The rule of thumb we used to estimate the minimum sample size is contested (Westland, 2010; Wolf et al., 2013). Per Soper (2025), our sample size would have been insufficient to detect small effects (effect size 0.1) but more than sufficient to detect

moderate effects (0.3). Thus it is plausible that predicted effects that were not supported by our findings were in fact small effects we lacked the power to detect. However given that the relationship between marginalised identity and obfuscation use is salient enough to have received media attention, it seems reasonable to expect a moderate effect size.

7.9 Conclusion

Obfuscation is ubiquitous on TikTok, driven primarily by a desire to post content that may otherwise be subject to censorship. Obfuscation seems to have become a part of the platform vernacular, and any attempts to improve user experience and user safety must take its widespread popularity into account. For some marginalised groups, such as Black (British) users, obfuscation provides an opportunity for creativity and play, whilst for others, for example queer TikTokers, use seems to be driven by a sense of necessity. Perceived censorship on TikTok patterns after that on other social media sites and in wider society, for example the more aggressive censorship of women's sexuality, and users may use obfuscation to "correct" perceived biases in the recommender and moderation algorithms that they believe will police their self-expression. Social media companies looking to improve the experiences of marginalised users must ensure their algorithms are fair and, crucially, *perceived to be fair*; failure to do so means user behaviour will continue to undermine safety features on the platform, such as TikTok's suicide support system, in their attempts to avoid algorithmic censorship. Users' perception of being "at-odds" with the moderation system additionally risks reputational damage, like TikTok's "endless cycle of censorship and mistakes" (Ohlheiser, 2021). Future work is needed to establish which fairness interventions would be most positively received by users, and how TikTok might encourage linguistic play, particularly valued by Black users, whilst also ensuring the efficacy of its safety mechanisms.

7.10 Learnings

This Chapter in particular exemplifies the benefits of the third maxim: a human-centric approach focuses on the impact of technology on people in context – and how they respond. By considering how the public interact with these models, we realise that the impact of the biased censorship system is mediated by user behaviour. Studying bias as a feature of the moderation and recommender systems in isolation would fail to

capture the true impact of biased censorship.

This Chapter suggests that the relationship between a biased algorithm and the people that it impacts may become like a game of cat and mouse. Moderation, intended to minimise harm (particularly to marginalised people), leaves people (particularly marginalised people) feeling censored. As a result, the users employ obfuscation techniques, but this can also undermine the moderation system's ability to prevent harm. This may drive platform owners to develop more stringent moderation, which in turn harms more marginalised users and encourages them to adopt more creative avoidance techniques. To end this game, and actually prevent harm (as well as align the platform with their commercial goals), platform owners must take into account perceived bias and more generally work with the affected communities.

This Chapter also demonstrates other benefits of my human-centric approach. Social science theory related to identity and language fruitfully inspired us to explore the role of community in obfuscation use. Likewise this Chapter demonstrates the benefits of a trans-disciplinary approach that is informed by the work of wider stakeholders including journalists and platform users.

Chapter 8

Conclusion

Inspired by work by scholars such as Su Lin Blodgett, Hanna Wallach and Kate Crawford (Blodgett et al., 2020; Wallach, 2014; Crawford, 2017), this thesis sought to improve the way that social bias evaluation is conducted in NLP. I have primarily argued for and demonstrated the benefits of a human-centric approach to studying social bias in NLP with five clear maxims, maxims shaped by ethics, social science and HCI research, and also by my own early work. These are: see NLP tools as part of large socio-technical systems; consider many sources of bias; focus on the impact of technology on people in context - and how they respond; be driven by social science theory and community knowledge; and address a broad range of demographics. I will begin by demonstrating the rich insights that can be gained from such a focus on human experiences by discussing two key findings that I have derived from my own work. I will then discuss the wider benefits of my human-centric approach to the field of NLP. Next, I will discuss the wider implications of my thesis, for those who adopt, regulate and report on NLP technologies. I will conclude with brief final thoughts.

8.1 Key Findings

In order to motivate my approach, I have, in addition to naming common limitations of contemporary work, sought to evidence the need for a human-centric approach through four "case studies". Herein I will discuss the valuable insights I gained by adopting this approach. To briefly summarise the findings of each of my "case studies", in Chapter 4 I showed that "low tech" NLP models can provided needed control, and that superficial debiasing leaves less salient groups impacted by harm. In Chapter 5 I showed that heuristic approaches to bias mitigation are widely rejected by the non-

cisgender community, and discussed how "obvious" solutions to bias can still lead to harm. In Chapters 6 and 7 I demonstrated that the felt impact of an algorithm is the result of human-AI interaction, and that the public will develop their own "solutions" to perceived bias. Whilst my work has addressed three very different tasks – sentiment analysis, TTI generation and social media content moderation – there have been commonalities in my findings (as well as, of course, in my approach), and overall two major findings from my thesis are that trying to prevent bias can cause harm, and that the public form complex beliefs about algorithms.

Considering the first key finding, for sentiment analysis systems (Chapter 4), this looked like ignore lists resulting in less effective models, particularly for queer content. For TTI models (Chapter 5), this looked like erasing certain identities. Further, our survey suggested that many heuristic attempts to prevent bias are widely rejected by the community as potentially harmful e.g. associating queer identities with warning labels. These techniques may give the appearance of less model bias, but do little to address harm. And if social bias is intended as a proxy for harm – why else would it be worth measuring? – it is a poor proxy where addressing the one actually worsens the other.

Often, bias mitigation techniques are used to make products palatable to a mass audience, at a low cost: the heuristic techniques discussed in Chapter 5 (and so widely rejected) were all inspired by techniques I have seen major tech companies employ to this end. Ignoring salient queer identity terms prevents the problem of bad PR, such as Thompson (2017), and is a much cheaper fix than say finetuning a model on more balanced data. For social media moderation (Chapters 6 and 7), this looked like seemingly silencing marginalised users in an attempt to prevent controversial content. Indeed, a recurring theme in this thesis has been the role of commercial priorities.

Unfortunately, commercial priorities have likewise influenced academic research in AI (Abdalla and Abdalla, 2021), and it seems plausible this has limited the rigour of bias measurement, even subconsciously. In this thesis, I have critiqued the tendency for bias research to consider only an abstract conceptualisation of bias, affecting a small set of identities, without reference to social science theory or community knowledge. Elsewhere I have discussed the issue of multi-lingual models being tested only for bias in English, or researchers employing only one of a battery of tests available (Goldfarb-Tarrant et al., 2023). Deliberately or not, the current approach to bias evaluation is very superficial, finding fault – but not too much fault. Going beyond this typical approach requires significant (financial) resources in terms of establishing mul-

tidisciplinary research teams (Tobi and Kampen, 2018; Fischer et al., 2011), and time spent conducting theory-grounded analyses and working with affected communities. It may be that practitioners are insufficiently resourced or perhaps motivated to conduct more thorough bias evaluation.

My second key finding is that affected communities form complex beliefs about the workings of these algorithms, evidenced in Chapters 5 through 7, particularly in free text answers. It is these beliefs that guide how users respond to AI, such that understanding them is vital to understanding the impact of a given model. Studies of what factors impact perceived fairness in automated decision making are numerous (Starke et al., 2022); I call for more research on what impacts perceived bias in other forms of AI such as generative AI. These complex beliefs extend to theories on how to mitigate harm. Suggestions for possible harms and how to improve were plentiful in Chapter 5, and the public are productive in their use of language to avoid biased censorship as evidenced in Chapter 7. This demonstrates significant creativity in people's interactions with AI, and whilst this allows them to counteract bias it may also result in their getting around safety guardrails, intentionally or not. A refrain I have often used when discussing my thesis is that humanity is infinitely more creative than a single team of developers, such that any attempt to prevent bias will always fall horribly short of possible use cases.

8.2 Benefits of my Human-Centric Approach to Social Bias Evaluation in NLP

Ultimately, my proposed approach will enable practitioners to better capture what they set out to capture! Specifically, adopting my five proposed maxims and centring humans in social bias research will allow practitioners to develop more valid bias evaluation methodologies, which will ultimately lead to more effective bias mitigation approaches. Thus this work is relevant to all those engaging in social bias research in the widest sense, be it within academic, industry or third-sector organisations. Whether motivated by a desire for social change or a desire to avoid expensive PR crises (Ohlheiser, 2021), practitioners will likely have an interest in improving the validity and effectiveness of their social bias research. This pertains not just to individual practitioners but also to the companies that employ them. My maxims can form the basis of an individual practitioner's personal work philosophy (as they have formed

my own), or can be used as the foundation for an entire company's approach to bias evaluation and mitigation.

Bias in abstract is often a poor proxy for the impact of a system on affected people. Beyond problematising the relationship between bias and harm (or rather, echoing previous work that has problematised this relationship such as Blodgett et al. (2020, 2021); Goldfarb-Tarrant et al. (2021); Cao et al. (2022) *i.a.*), I have shown what valuable insight can be gained from directly investigating people's experiences of harm. I was able to identify potential harms of TTI models that would be hard to detect in model output alone (such as weaponisation and the enforcement of gendered beauty standards). I was able to measure marginalised TikTok user's experiences of discrimination, even where considering reported removal rates alone would not reveal disparities. Further, studying the affected populations allowed me to understand what they are currently doing to handle model bias, be it non-use (as I found for TTI models), or use of obfuscated language on TikTok. Understanding user behaviour is vital to developing effective mitigation strategies. For example, TikTok will now have to work within the culture of creative language use they have engendered on the platform to tackle the issue of perceived discrimination, whilst maintaining their desired level of safety – of sanitisation – on the platform.

In addition to questioning the relationship between conventional bias metrics and harm, I hope my work and proposed approach will encourage practitioners to interrogate the other normative decisions they make when evaluating harm (which I have discussed in more depth elsewhere (Goldfarb-Tarrant et al., 2023)). That is to say, how their values are embedded in the design of the system (Friedman et al., 2013; Talat et al., 2021). Every decision made from conceptualisation through to delivery can introduce bias into the system, and thus potentially elicit harm, and I encourage practitioners to be aware of the power they yield when making normative decisions such as how to conceptualise bias, what good model behaviour looks like etc.

My work has shown the value of qualitative data to measuring social bias in NLP. My systematic analysis of survey responses and interview data in Chapter 5 allowed me to identify novel harms and potential solutions. There is a tendency to rely on quantification to legitimise ethics research (Widder et al., 2023), but there is a lot of value in looking at insights from the community, that might not be systematically analysed,[1] but are nonetheless vital. Much can be gained from giving the community a

[1] Whilst I only conduct systematic qualitative analysis in Chapter 5, I believe there is also great value in looking at free text answers in detail, as I do across Chapters 5 through 7.

chance to share. During my PhD I completed an internship at Google Research investigating the harms that large language models (LLMs) might do harm to the trans and nonbinary community. We surveyed community members and conducted multiple workshops, which allowed us to create an incredibly rich and granular taxonomy of possible harms. Centring community experiences allowed us to break down a complex and conceptually broad concept such as "transphobia" into specific harmful model behaviours, allowing us to propose heuristic evaluation metrics with strong validity, which will hopefully ultimately lead to targeted mitigation (Ungless et al., 2025a).

Of course, centring human experiences has value to the field of NLP well beyond social bias research. Considering the developers behind new AI technologies invites us to reflect on their values (Birhane et al., 2022a). It also invites us to consider how developers' commercial priorities might influence their work (Abdalla and Abdalla, 2021). By making the socio- part of socio-technical systems more salient, we can ensure that improvements to the system are genuinely effective and achieve algorithmic reparation (Davis et al., 2021). Elsewhere, I create a taxonomy of risks for TTI models focused on the impact of these technologies across a very broad range of stakeholders, including image subjects and those with limited access to verification tools (Bird et al., 2023). By making these stakeholders explicit, I make them more salient in the field, and in doing so hopefully ensure that future attempts at risk mitigation and regulation take the needs of these stakeholders into account.

8.3 Wider Impact

The ramifications for my proposed approach to social bias research extend beyond the field of NLP or even computer science. Most obviously, my critique of current social bias evaluation practices is relevant to those wishing to integrate NLP technologies into their own services and products. The recommendations I make at the end of each of my "case studies" (Chapters 4 - 7) are rendered tangible by my focus on specific use contexts. However, as I argued throughout Chapter 2, bias evaluation often happens without consideration for specific use cases. This suits the purposes of those promoting "general purpose" language technologies (they need not engage in the work of defining the risks of particular end uses, which also helps to maintain their distance from accountability), but this also means that their evaluations have limited value. Those adopting third-party NLP technologies should be aware that bias, fairness, harm and risk evaluations were likely carried out in some imagined "neutral" context, and that

before deploying the model it is vital they perform their own harm evaluations.

My call to focus on harms in context rather than bias in abstract also has ramifications for regulators. During the completion of my PhD, the EU AI Act was published, and I commend the focus on specific use contexts.[2] Regulators should resist the excitement of high-powered general purpose models (also known as "foundation models" – though many resist this term (Venté, 2023)) and "artificial general intelligence" (AGI) and continue to develop tangible regulations for specific use cases. Specifically, I recommend that all so-called general purpose models be evaluated for harm in the most likely and most high risk possible use cases, and must meet acceptable standards before release (even limited release). Data sets such as Wildchat (Zhao et al., 2024), which contains records of real user conversations with ChatGPT, can guide regulators in establishing likely use cases; the EU AI Act offers a strong foundation for establishing high risk use cases. Plausible harms include risks to privacy, to personal wellbeing and safety. Where regulators seek to evaluate risks of discrimination and stereotyping due to social bias, they might be guided by my proposed maxims to develop robust and valid tests, in collaboration with independent social bias practitioners.

I would also caution policy makers to be mindful that the felt impact of NLP technologies or wider AI will be determined not just by the algorithms but by the public's beliefs about those algorithms, as I explore in Chapters 5 through 7. Thus policy makers should consider both the models themselves and the way that companies communicate about these models – and also be aware of the limits of these regulatory attempts! We might hope to regulate algorithms, and science communication; we cannot hope to regulate people's felt experiences.

Throughout this thesis I have drawn on the work of tech journalists who have amplified the voices of those speaking out against experiences of automated discrimination. I anticipate journalists will continue to play a pivotal role in holding the NLP industry accountable, at least in the "court of public opinion", and I hope that my five principle issues and five proposed maxims are presented with sufficient clarity to be adopted as a foundation for further critique. Journalists can also play a role in making explicit the normative decision that developers have made whilst designing their products (which I have highlighted as a potential source of bias), allowing the public to question whether those norms align with their own values.

[2] I was disappointed to discover that the EU AI Act stops short of attempting to regulate the use of AI in military contexts – surely the most high risk of all.

8.4 Final Thoughts

Social bias is used as a proxy for measuring the harms of a model. But even when measured accurately, bias is often a poor proxy for harm. As I and others have shown, existing bias measurement methods for NLP technologies, like language models, frequently have poor validity and reliability. That is to say, they do not measure what they claim to measure in a consistent way. Coupled with the fact that these models exist as part of socio-technical systems in which stakeholders can introduce their own biases, this means measuring bias upstream and in abstract seems a fruitless exercise. It is only within specific use contexts that we can understand the negative impact of these models and collaborate with those impacted to develop meaningful solutions. Although the title of my thesis makes reference to social bias research, I ultimately argue that measuring bias (in abstract) is pointless, and we should refocus our efforts on measuring harms in context. I am far from the first to raise these critiques; I can only hope that adding my voice to the chorus will amplify the message, and enrich it with five clear maxims that improve the validity of social *harm* research. My work also progresses our understanding of queerphobia in NLP, with insights into how sentiment analysis, image generation and social media platform algorithms can harm the queer community – and how the community is pushing back!

Appendix A

Chapter 3: Additional Terms

In Chapter 3 we focus our discussion on the selection of terms related to our primary hypotheses (diverse queer terms, terms across the gender spectrum, terms related to trans status and terms specific to ethnic communities). In the following we provide additional discussion of identity terms and combinations present in the data set unrelated to these primary hypotheses.

The terms list includes multiple plurisexual identities (those attracted to multiple genders), namely "pansexual, fluid" in addition to "bisexual" from (Dixon et al., 2018). Plurisexual individuals often face discrimination from both queer and non-queer people, known as double discrimination (Mereish et al., 2017). The data set therefor allows for the detection of bias against plurisexual individuals (as well as against the individual sexualities). Reiterating a point we make above, when no sexuality is given, the norm (monosexuality) will be assumed. Note that whilst "queer" is widely used by plurisexual people, it is also used by monosexual people (Kolker et al., 2020), so we do not consider it plurisexual. Note "fluid" can refer to fluidity of gender but we only included this in the sexuality category. We also include in the sexuality list two ace-spectrum identities (Hille et al., 2020), namely "asexual, demisexual", which will enable exploration of ace-phobic bias, which is often overlooked in research on queer-phobic bias (Galop, 2021).

In order that the data set be useful for detecting bias against plurisexual identities, we combined terms relating to sexuality presentation style (e.g. "butch"), where monosexuality is assumed, with the term "bisexual". Our research at the time suggested "stud" was used by black women who exclusively love women (Lane-Steele, 2011), and as "bisexual" is a non-ethnicity specific term whilst "stud" is ethnicity specific, we felt it would be inappropriate to combine the terms. However, there has been

a movement on social media bringing attention to bisexual studs.[1] We welcome researchers to use our code to generate sentences including "bisexual stud", and further welcome inclusion of more ethnicity specific terms in our data set.

When constructing the data set and in our analyses we treat "queer, LGBT" and "LGBTQ" as sexuality terms. We felt this was the most appropriate way to include these umbrella terms without introducing a fourth identity term category (and thus effectively quadrupling the size of the data set). Although they are intended to encompass all queer identities, it is our intuition that the cisgender norm is often assumed despite the inclusion of T for transgender in the two acronyms. Indeed, LGBT(Q) is often used in direct opposition to straight, despite transgender straight people being part of the LGBT community (see for example Zane (2019)). Another researcher might prefer to exclude variations with LGBT and LGBTQ when considering non-transgender identities (we treat these as "assumed cisgender" identities).

[1] https://www.tiktok.com/@ainabreiyon/video/7058039963786431791?lang=en

Appendix B

Chapter 4: Survey

B.1 Demographic Information

Q: What is your age? Please answer in years.

Responses: range 19-57, mode 25, mean 30.

Q1: What is your gender identity?

Options: male, female, nonbinary, genderqueer, third-gender, genderfluid, gender non-conforming, pangender, two-spirit, agender, questioning, prefer not to answer, other.

Note: A transcription error resulted in the options "male, female" in place of "man, woman" from Dev et al. (2021). Typically "male, female" are more associated with "biological sex" than "man, woman" which may have influenced respondents' answers, although the question explicitly asked about gender.

Responses given in Table B.1.

Q2: What is your sexual orientation?

Options: lesbian, gay, bisexual, asexual, pansexual, queer, straight, questioning, prefer not to answer, other.

Responses given in Table B.2

Q3: What pronouns do you use?

Options: he/him, they/them, she/her, xe/xem, e/em, ze/hir, any pronouns, I don't use pronouns, I am questioning my pronouns, prefer not to answer, other.

Responses given in Table B.3

Q4: Are you trans?

Options: yes, no, I am questioning my gender, prefer not to answer. In retrospect, it may have been more appropriate to give the option "I am questioning my trans status".

Gender	%of total responses	Count
Male	2.9%	1
Female	22.9%	8
Nonbinary	71.4%	25
Genderqueer	20%	7
Genderfluid	8.6%	3
Gender non-conforming	14.3%	5
Agender	17.1%	6
Questioning	11.4%	4
Prefer not to answer	2.9%	1
Other – "trans"	2.9%	1
Other – "I'm also intersex"	2.9%	1
Other – "Woman"	2.9%	1

Table B.1: Table of selected gender identities. Respondents could select multiple gender terms.

Responses given in Table B.4

Q5: In a few words, how would you describe your ethnicity? Options: text response

The majority of respondents (26) described themselves as explicitly white or Caucasian. Four named a European origin (none of these identified as Black, Latinx and/or Indigenous or as a person of colour); as white is the norm in Europe (Kantola et al., 2022), this suggests 30 of our 35 participants are white/Caucasian.

Q6: Are you Black, Latinx and/or Indigenous ?

Options: yes, no, prefer not to answer.

Responses given in Table B.5

Q7: Are you a person of color?

Options: yes, no, prefer not to answer.

Notes: Not all respondents who identified as Black, Latinx and/or Indigenous also identified as a person of colour and vice versa. Please note the discussion in Section 5.9 about use of the term "latinx".

Responses given in Table B.6

Q8: What is/are your native language(s)?

Options: text response

Sexual orientation	% of total responses	Count
Lesbian	17.1%	6
Gay	8.6%	3
Bisexual	34.3%	12
Asexual	5.7%	2
Pansexual	17.1%	6
Queer	42.9%	15
Straight	2.9%	1
Prefer not to answer	2.9%	1
Other – "i try not to label myself "	2.9%	1
Other – "Bottom"	2.9%	1

Table B.2: Table of selected sexual orientations. Respondents could select multiple terms.

The vast majority of participants (27) had English as a native language. Other native languages include German, French, and BSL.

Q9: Which country do you live in now?

Options: text response

Responses are summarised in Table B.7. All bar one of our respondents (34) are from Western countries, namely North America, Europe or Australia.

Q10: Briefly, how would you describe your occupation?

Options: text response

Ten respondents described themselves as students. The next most common occupation was software engineer. Other occupations include photographer, creative professional, UX designer and therapist, suggesting we were able to capture the diverse perspectives of those working outside the field but with an interest in AI.

Q11: Briefly, how would you describe your familiarity with AI?

Options: text response

The majority of respondents referenced work or education as being the source of their familiarity, though some named an interest in the topic for example as a "science magazine reader". One respondent answered none but rated themselves as 2/5 in terms of familiarity with AI.

Q12: How would you rate your familiarity with AI?

Options: Likert scale 1-5 from "Very little knowledge" to "Expertise (I work in

Pronoun set	% of total responses	Count
He/him	17.1%	6
They/them	68.6%	24
She/her	34.3%	12
E/em	2.9%	1
Any pronouns	11.4%	4
I am questioning my pronouns	17.1%	6
Other - “Elle/le”	2.9%	1
Other - “Ey/Em”	2.9%	1
Other - “xey/xem”	2.9%	1
Other - “fae/faer”	2.9%	1

Table B.3: Table of selected pronouns. Respondents could select multiple terms.

Response	% of total responses	Count
Yes	85.7%	30
No	2.9%	1
I am questioning my gender	5.7%	2
Prefer not to answer	5.7%	2

Table B.4: Table of responses about trans status.

Response	% of total responses	Count
Yes	8.6%	3
No	91.4%	32

Table B.5: Table of responses to question about identifying as Black, Latinx and/or Indigenous.

Response	% of total responses	Count
Yes	8.6%	3
No	91.4%	32

Table B.6: Table of responses to question about identifying as a person of colour.

Region	% of total responses	Count
US	31.4%	11
UK	34.3%	12
Europe excl. UK	22.9%	8
Canada	5.7%	2
Australia	2.9%	1
Colombia	2.9%	1

Table B.7: Table of responses to question about current country of residence.

AI)".

Responses are summarised in Figure B.1. All respondents considered themselves to have greater than "very little knowledge". The mean rating was 3.8.

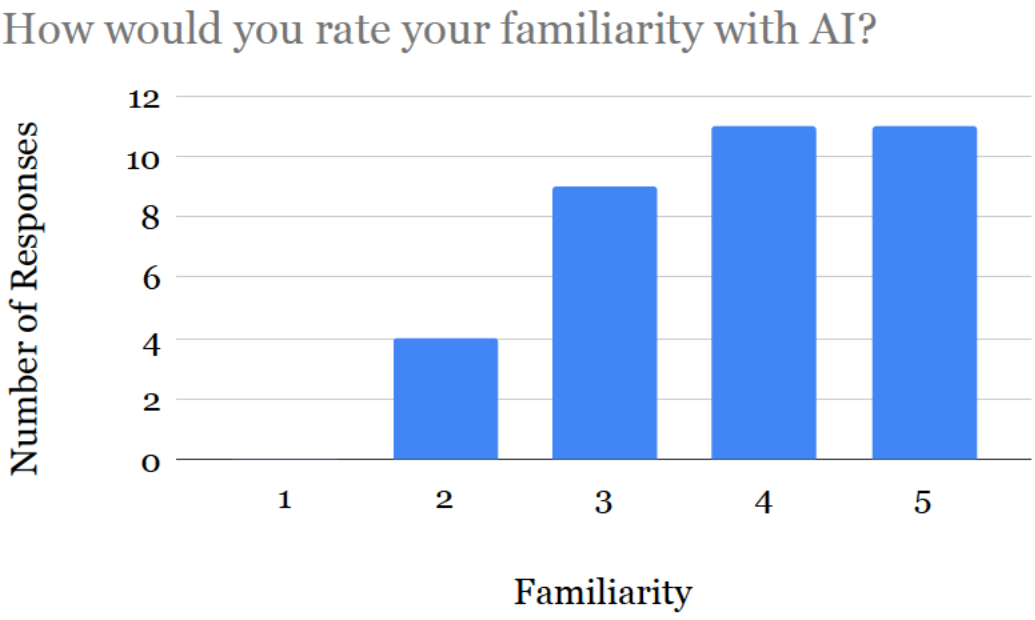


Figure B.1: Count of responses for each familiarity rating.

B.2 Potential for Harm

Q13: Have you tried out one of these systems before, including during this survey?

Options: yes, no

The vast majority of respondents (28) answered yes.

Context	% of total responses	Count
Education	91.4%	32
Art/creativity	85.7%	30
Marketing	94.3%	33
Architecture/ real estate/ design	37.1%	13
Research	71.4%	25

Table B.8: Table of responses to question 15 about contexts in which harm might occur to non-cisgender individuals.

Q14: Can you think of scenarios where use of text-to-image models could have undesirable outcomes for non-cisgender people, due to their application in the above or other use cases?

Options: yes, no

The overwhelming majority of respondents (33) answered yes.

Q15: Please select in which of these use cases harms might occur.

Options: education, arts/creativity, marketing, architecture/ real estate/ design, research, other.

Notes: These options are derived from DALL·E 2 documentation detailing possible future commercial use of the model. A flaw in the study design meant this question was mandatory even for those who answered "no" to the previous question. Of the two participants who answered no, one wrote "none" in the "other" option and the other selected Education, but neither provided a description of a scenario (below).

Responses are summarised in Table B.8. Two respondents provided "other" contexts of use – one referenced religious and political channels, and the philosophical, psychological and sociological fields, and the other wrote that they were concerned about the "reinforcement of heteronomativity in any context". The majority of respondents could imagine harm in each of the contexts except "Architecture/ real estate/ design". In particular respondents were concerned about "Marketing", "Education" and "Art/creativity" (over 3/4 of respondents felt harm might occur in these contexts).

Q16: Please select how severe you think these harms might be.

Options: Likert scale 1-5 from "No impact on lives" to "Significantly hinders lives".

Responses are summarised in Figure B.2. The average rating was 3.3. Almost

all participants (32) felt that the harms would have some impact on non-cisgender individuals' lives.

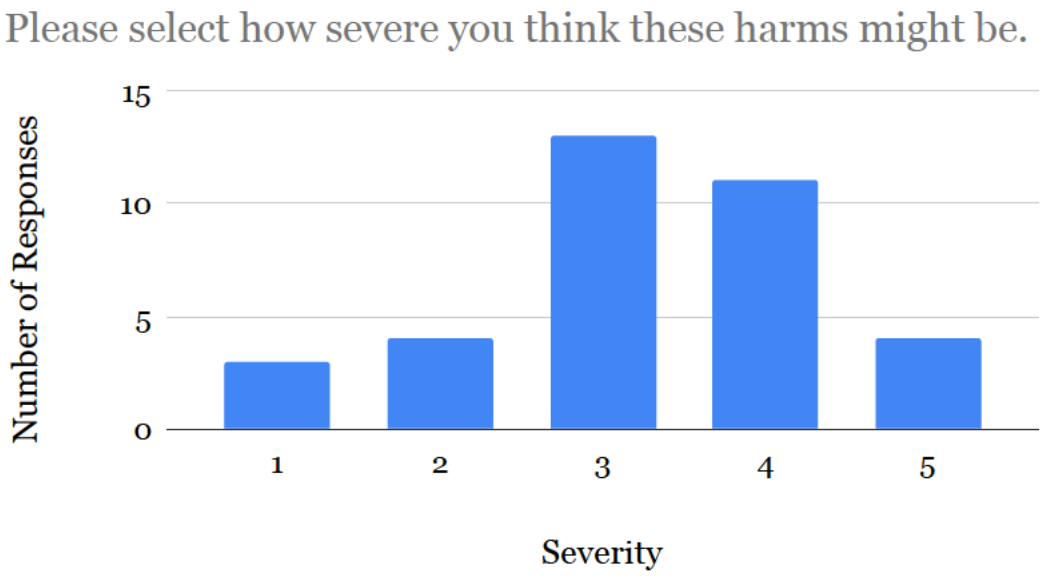


Figure B.2: Count of responses for each severity rating.

Q17: Please describe a specific scenario(s) where harm might occur against non-cisgender people.

Options: text response

In contrast to Dev et al. (2021) we do not ask survey participants to distinguish between representational and allocational harms, in order to reduce their work load. We label which category of harm they describe, whether it relates to how a group is represented or which services a group has access to, or both. We also identify which use cases are relevant to the harm they mention, again to reduce work load. Using a deductive-inductive approach, we also develop codes and establish themes based on the responses. One author was lead coder, developing the codebook of 17 codes. Both the lead coder and a second author applied this codebook to the responses. The coders refined the codebook through discussion, leading to a final inter-coder reliability of $\kappa = 0.74$. Themes were identified by the lead coder and discussed and finalised through discussion between all authors.

Loosely reflecting the responses to Q15, the contexts of use mentioned by respondents were education, art/creativity, marketing, and less frequently research. A high number of representational harms were identified, and very few allocational harms.

A prominent theme was the potential impact on real world behaviours and beliefs

that content produced by the models might have. Frequently, respondents spoke of the output not just reflecting but *reinforcing* stereotypes and prejudices. Some felt the tools could create new beauty standards and lead to emotional harm.

Several respondents expressed concern about intentionally abusive use of these systems. They felt they might be used to create propaganda or transphobic material, or the training data needed to create a trans recognition system. These examples far outnumbered explicit references to unintentional harms.

A number of respondents explicitly referenced the role that training data played in bringing about harm, reflecting the knowledge of our respondents.

B.3 Proposed Solutions

Respondents were asked to rate on a likert scale of 1-7 ("Extremely dissatisfied (I would not like to see this solution implemented)" to "Extremely satisfied (I would like to see this solution implemented)"). A rating of 4 indicates neither satisfied or dissatisfied. They were also invited to optionally respond to the question "Can you foresee any potential harms or benefits to this solution?" for each one.

Solution 1: The model generates an image based on the text (no change to current behaviour.)

Responses are given in Figure B.3. Most respondents were unsatisfied with this "solution" (to change nothing), with a mode of 3 and a mean of 3.5 (both below 4). However the spread of responses indicates this is not universally disliked. Text responses in particular highlighted concerns about stereotyping,

Solution 2: The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text.

Responses are given in Figure B.4. This solution was the least popular, with a mode of 1 and a mean of 2. No respondents were clearly satisfied with this solution. Many respondents wrote this would lead to erasure and othering. A respondent identified it would be hard to "keep up" with queer slang, or handle ambiguous words.

A simple heuristic like ignoring minority identity terms to avoid producing stereotyped content is clearly not satisfactory to the community.

Solution 3: The model generates an image based on the text but includes a warning that the output might be offensive.

Responses are given in Figure B.5. This solution was also fairly unpopular, with a mode of 2 and a mean of 3.0, although the bimodal results suggest some users would

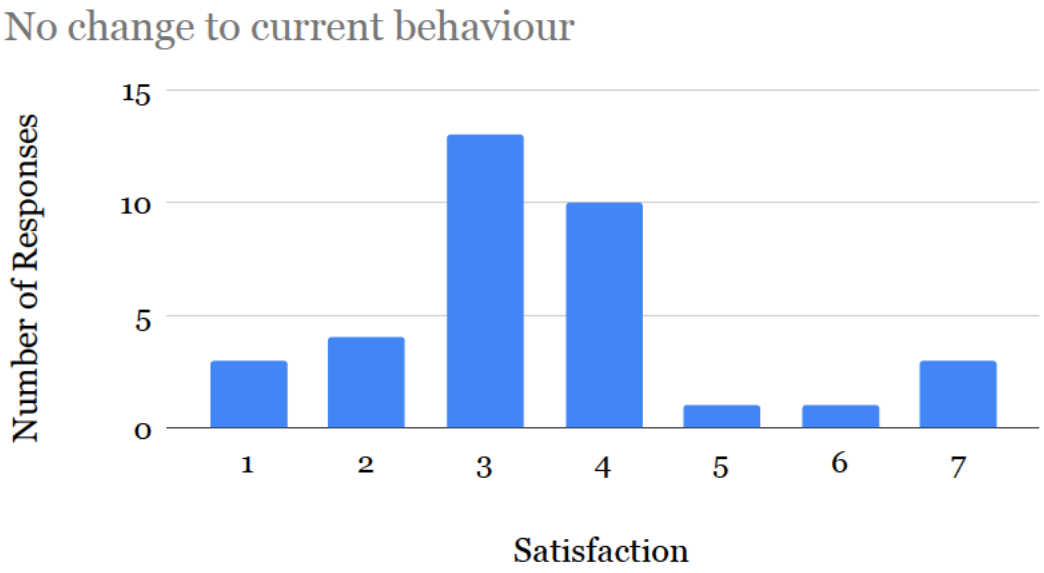


Figure B.3: Count of responses for each satisfaction rating for Solution 1.

be slightly satisfied by this solution. Several respondents expressed that they felt this was not a "real" solution to the issue. Some felt strongly that appending this warning to every image suggested transness itself was offensive. However, as suggested by the second "peak", some respondents felt a warning offered an okay interim solution.

Solution 4: The model ignores all gender identity terms in the text input and generates an image based on the rest of the text.

Responses are given in Figure B.6. This solution was very unpopular, though less so than ignoring only non-cisgender identity terms, with a mode of 1 and a mean of 2.5. Some respondents expressed concern about the model "defaulting" to represent only a single gender rather than diverse results. Respondents again mentioned erasure. Several respondents mentioned compromised functionality. Some felt it would be difficult to implement.

Solution 5: The model is trained on additional images containing non-cisgender individuals, so it better learns to generate images of non-cisgender people.

Responses are given in Figure B.7. This solution was by far the most popular, with a mean of 5.3 and a mode of 7. However, as Figure B.7 demonstrates, this solution is not universally popular, and in text responses respondents expressed concern about the challenge of gathering truly representative data, and the risk of reinforcing stereotypes. Some expressed concern about the risks of gathering images of marginalised people.

Solution 6: The model effectively ignores the non-cisgender identity terms in the

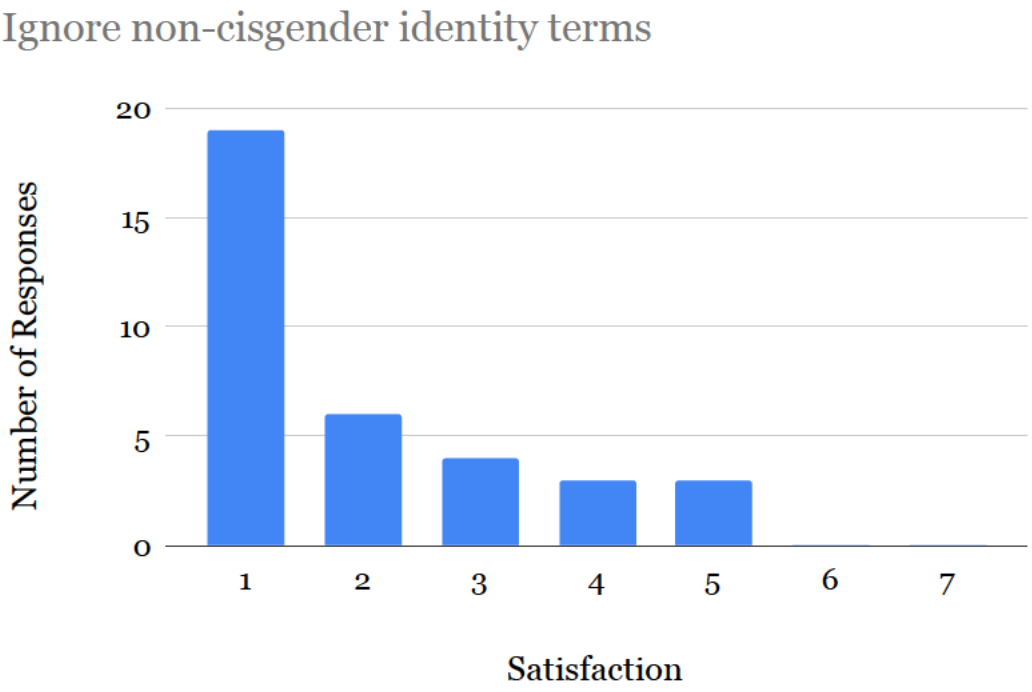


Figure B.4: Count of responses for each satisfaction rating for Solution 2.

text input and generates an image based on the rest of the text, but a flag or pin or symbol is used to indicate gender diversity.

Responses are given in Figure B.8. This solution had a mode of 1 and a mean of 2.8, suggesting it was largely unpopular (though a small number were satisfied with this solution). Some respondents expressed that this solution had potential, because it no longer required using how a person looks to capture their identity. Others felt it was a "cop out", and some were concerned about the othering or stigmatising effect of explicitly labelling queer individuals.

Solution 7: The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, with a warning that to avoid harmful misrepresentation the model ignores non-cisgender identity terms.

Responses are given in Figure B.9. This solution was largely but not universally unpopular, with a mode of 1 and a mean of 2.7. Respondents expressed a preference for ignoring the terms alongside an explicit warning over simply ignoring the terms in their text responses, but many argued the same issues of erasure and compromised functionality were at play. A few saw it as a short-term solution, but many argued it was again a "cop out".

Other Solutions

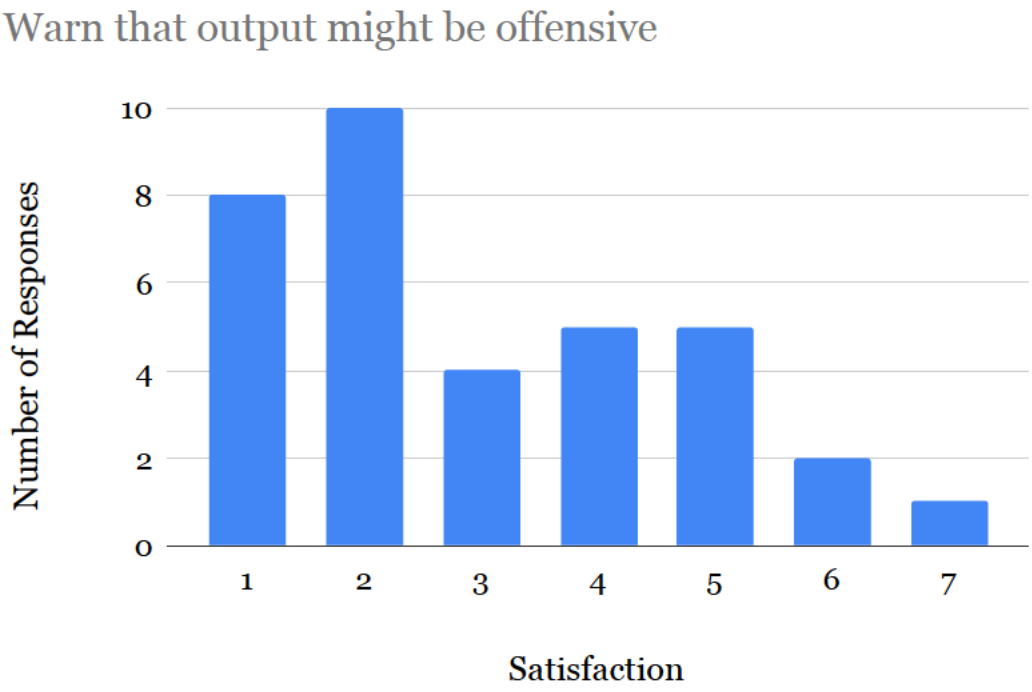


Figure B.5: Count of responses for each severity rating for Solution 3.

Respondents were then asked "Can you think of any other solutions to how models should handle non-cisgender identities? (Optional)". The majority (22) of respondents provided their thoughts. We conducted a qualitative analysis of these answers, using an inductive approach. One author developed the codebook of 22 codes using a "bottom-up" approach (driven by the data), which was then applied to the responses by a second author to establish inter-coder reliability, as a measure of code reliability. The lead coder established themes based on these codes and these themes were discussed and finalised between the authors. The major themes we established were the need for representative data; unhappiness with the proposed heuristics; the necessity of wider changes; community involvement; a desired ability to customise images.

One theme we established was the need for representative training data, echoing the most popular proposed solutions. Many respondents emphasised the need for additional data, others focused on the need to curate the training data to ensure "a diverse and representative set of images" (white, queer, nonbinary + gender nonconforming, 23).

A second theme that emerged was that of unhappiness with the proposes heuristics, with respondents seeing these as outright unsuitable or suitable only as temporary solutions.

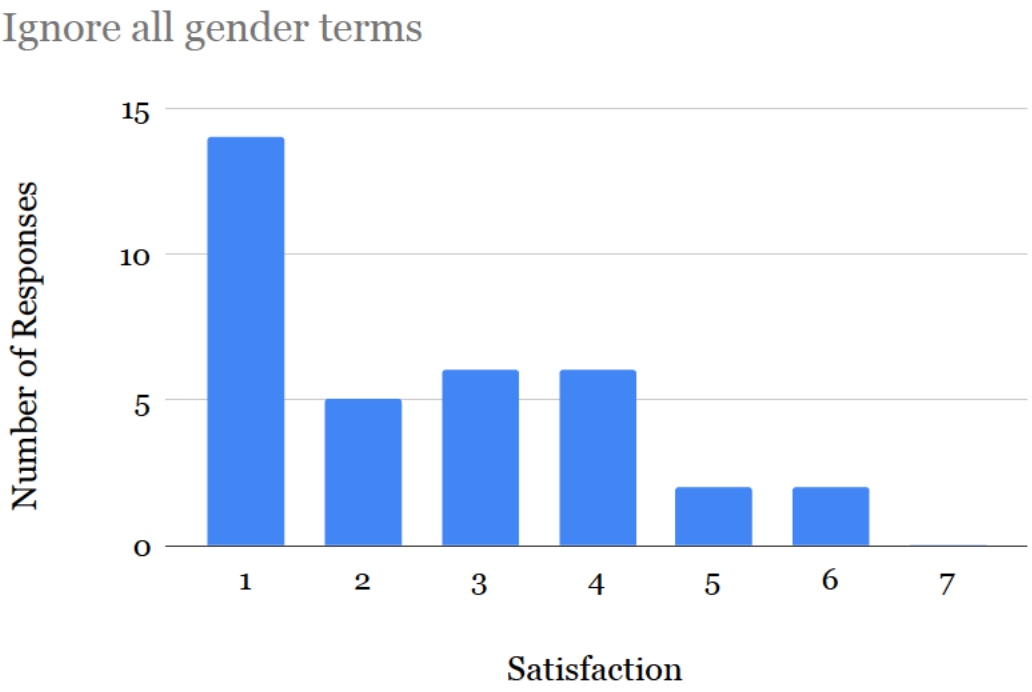


Figure B.6: Count of responses for each satisfaction rating for Solution 4.

A broad theme in the responses was the need for wider changes, encompassing both extensive changes to the model, as well as societal changes – "may require uhhh fixing society generally" (white, bisexual, genderqueer + questioning, 30). Respondents all mentioned the need to improve outcomes for other marginalised identities.

Another theme to emerge was the need for community involvement – respondents discussed the general need for non-cisgender people to be involved in the development of such models, and two suggested involving non-cisgender individuals as part of a reinforcement learning approach to improve the models' representation of the community.

The final theme represents a novel solution, which is to allow for post-hoc modification of the generated images. This would mean users could tweak the gender presentation and/or include symbols and pins to signify identity, reminiscent of one of the proposed heuristics but with greater user agency.

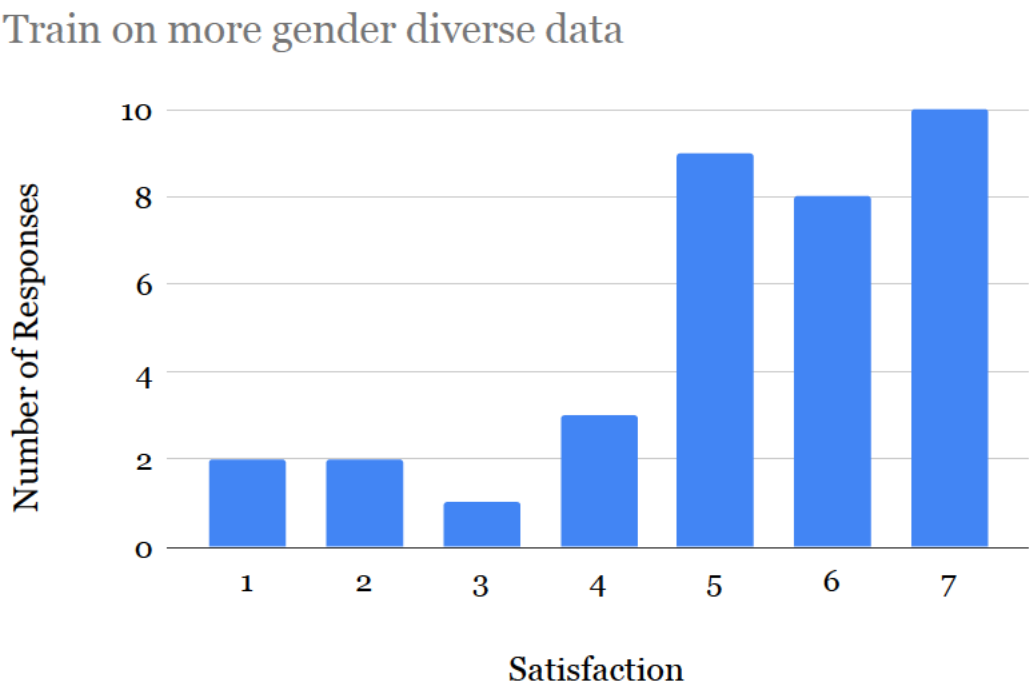


Figure B.7: Count of responses for each satisfaction rating for Solution 5.

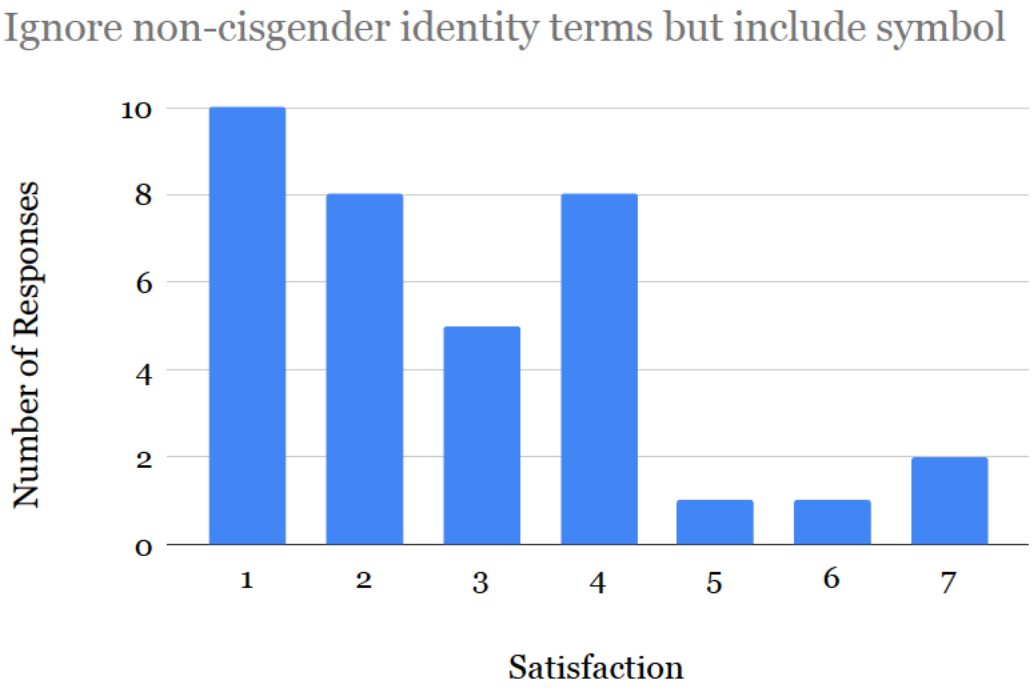


Figure B.8: Count of responses for each satisfaction rating for Solution 6.

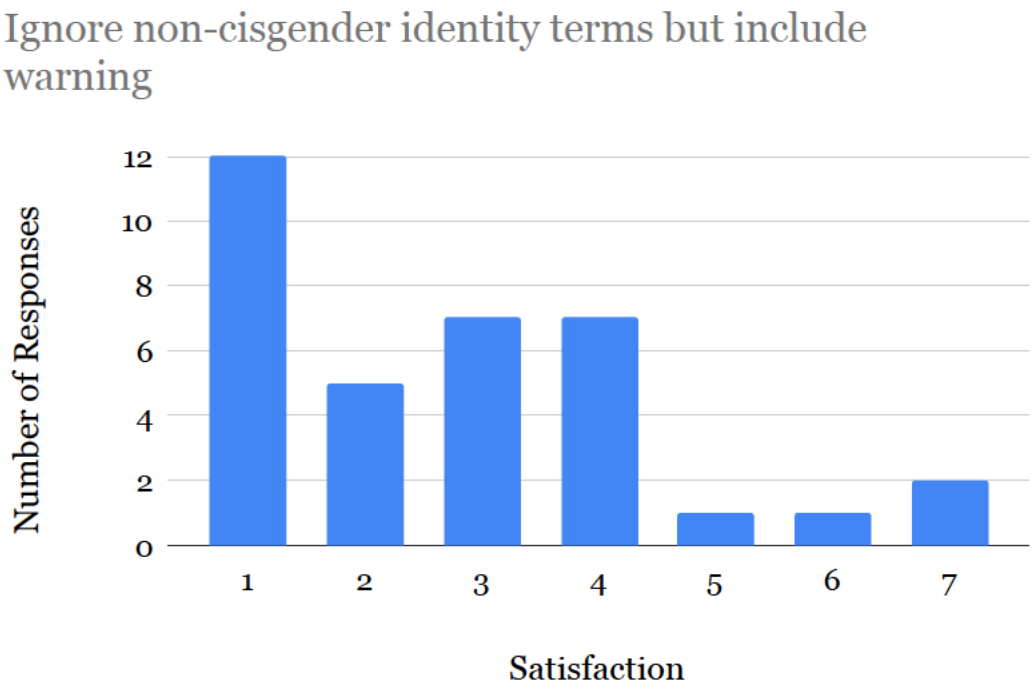


Figure B.9: Count of responses for each satisfaction rating for Solution 7.

Appendix C

Chapter 5: Supplementary Material

Consent

Participant Information Sheet

Project title:	TikTok Censorship
Principal investigator:	Björn Ross
Researcher collecting data:	Eddie Ungless, Nina Markl
Funder (if applicable):	UKRI

This study was certified according to the Informatics Research Ethics Process, RT number 747333. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?
Eddie Ungless, he/they, PhD candidate at the Centre for Doctoral Training in NLP, University of Edinburgh
Nina Markl, she/her, PhD candidate at the Centre for Doctoral Training in NLP, University of Edinburgh
Björn Ross, he/him, Lecturer in Computational Social Science at the Institute for Language, Computation and Cognition, University of Edinburgh

What is the purpose of the study?
The goal is to understand TikTok users' perception of being algorithmically censored, and their motivations for using techniques that help to avoiding censorship. The study will comprise of a survey.

Why have I been asked to take part?
You have been asked to take part because you have indicated that you are or have been a user of TikTok.

Do I have to take part?
No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?
This study is a survey involving some multiple choice and some free text questions. You will be asked to respond to questions about your use of TikTok. We will ask about the kinds of content you view and post.
You will be asked about your experience of algorithmic censorship. You will be asked about your experience of the techniques people use to avoid censorship. You will be asked how censorship makes you feel.
You will also be asked to provide demographic information including gender, sexuality, ethnicity, disability status, religion, socioeconomic status and political beliefs. We are gathering this sensitive information as we want to understand how users of different social identities behave on TikTok.

We are only recording your answers to these questions.

This survey should take around 14 minutes to complete.

Compensation.
You will be paid £1.70 (equivalent to £6/hr) for your participation in this study.

Are there any risks associated with taking part?
There are no significant risks associated with participation. Your answers will be anonymised so the risk of sensitive information becoming known is extremely low.

Are there any benefits associated with taking part?
No, other than contributing to our understanding of users' experiences of censorship on social media.

What will happen to the results of this study?
The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

Data protection and confidentiality.
Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a

unique participant number rather than by name. Your data will only be viewed by the research team Eddie Ungless, Nina Markl and Björn Ross.
All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) or held securely by Qualtrics, the questionnaire software provider.

What are my data protection rights?
The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.
For general information about how we use your data, go to: edin.ac/privacy-research

Who can I contact?
If you have any further questions about the study, please contact the lead researcher, Eddie Ungless,
If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.
If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

Consent
By proceeding with the study, I agree to all of the following statements:
- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations.
- I allow my data to be used in future ethically approved research.

Refuse and exit	Agree and continue
○	○

Use of TikTok

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

${e://Field/PROLIFIC_PID}

Do you use TikTok?

- ○ Yes I currently use TikTok
- ○ I previously used TikTok but am not a current user
- ○ I have never used TikTok

When did you first use TikTok?

- ○ More than three years ago
- ○ Between three and one years ago
- ○ Between 12 and 6 months ago
- ○ Between six and three months ago
- ○ Between three and one month ago
- ○ Less than one month ago

When did you stop using TikTok

- ○ More than three years ago
- ○ Between three and one years ago
- ○ Between 12 and 6 months ago
- ○ Between six and three months ago
- ○ Between three and one month ago
- ○ Less than one month ago

What were your reasons for stopping using TikTok?

- ☐ It was negatively impacting my quality of life
- ☐ The content was too negative
- ☐ To escape harassment / bullying
- ☐ A loved one asked me to stop using the app
- ☐ My account was deleted / banned
- ☐ Too much data use / phone storage use

- [ ] Too much moderation / censorship
- [ ] The amount of advertising
- [ ] Concerns about security / privacy
- [ ] To have more free time
- [ ] Not enough of my friends were using it
- [ ] The content wasn't entertaining anymore
- [ ] The format of the app / content
- [ ] [ ] Other

On average, how often did/do you use TikTok to view content?

- ○ I have never viewed content
- ○ 0-3 times a month
- ○ 1-2 times a week
- ○ Multiple times a week
- ○ 1-2 times a day
- ○ >3 times a day

On average, how often did/do you use TikTok to post content?

- ○ I have never posted content
- ○ 0-3 times a month
- ○ 1-2 times a week
- ○ Multiple times a week
- ○ 1-2 times a day
- ○ >3 times a day

On average, how long did/do you spend on TikTok on the days that you use(d) the app?

- ○ Less than 15 minutes a day
- ○ Between 15 and 30 minutes
- ○ Between 30 minutes and 1 hour
- ○ Between 1 and 2 hours
- ○ Between 2 and 4 hours
- ○ More than four hours a day

What type of content have you viewed on TikTok (select all that apply)

- [ ] Telling jokes
- [ ] Telling stories
- [ ] Completing sound trend/ challenge
- [ ] Poster making fun of themselves
- [ ] Prank
- [ ] Dance
- [ ] Learn something
- [ ] Beauty/makeup
- [ ] Lip Sync
- [ ] Singing (real voice)
- [ ] Exercise
- [ ] Expose someone
- [ ] Duet
- [ ] Show off body (e.g. flexing, revealing clothing)
- [ ] Sharing personal information
- [ ] Politics
- [ ] Tribute to recently deceased person
- [ ] Pets
- [ ] Other (please describe)

What type of content have you posted on TikTok (select all that apply)

- [ ] Telling jokes
- [ ] Telling stories
- [ ] Completing sound trend/ challenge
- [ ] Poster making fun of themselves
- [ ] Prank
- [ ] Dance
- [ ] Teach something
- [ ] Beauty/makeup
- [ ] Lip Sync
- [ ] Singing (real voice)
- [ ] Exercise
- [ ] Expose someone
- [ ] Duet
- [ ] Show off body (e.g. flexing, revealing clothing)
- [ ] Sharing personal information
- [ ] Politics
- [ ] Tribute to recently deceased person
- [ ] Pets
- [ ] Other (please describe)

Did/do you view content in English on TikTok?

- ○ Yes, exclusively in English
- ○ Yes, in addition to other languages
- ○ No

What languages do/did you view content on TikTok in? List in order of most to least content viewed in that language.

Did/do you post content in English on TikTok?

- ○ Yes, exclusively in English
- ○ Yes, in addition to other languages
- ○ No

What languages do/did you post content on TikTok in? List in order of most to least content posted in that language.

Please note for the remainder of this survey we are interested in your use of TikTok to view and post **English language** content.

Removal and suppression of content

We are going to ask you about your experience of algorithmic censorship by TikTok.

On TikTok, some users are concerned about algorithmic censorship, which refers to when a computer program or algorithm designed to moderate content removes or suppresses the content they post i.e. when a video gets very few views and you believe it is because it is being suppressed by the platform.

In this study, when we say **censorship** we mean both when content is **removed** and when content is **suppressed**.

Have you had content removed by TikTok?

- ○ No
- ○ Yes

Please select which of the types of content you have had removed by the platform.

I have had posts removed that contained (you can select multiple):

- [ ] Covid-related content
- [ ] Sex related content for an erotic purpose
- [ ] Content some may find offensive or inappropriate
- [ ] Self-referential use of slur i.e. d*ke by a lesbian
- [ ] Content insulting or criticizing dominant group (e.g., men, white people)
- [ ] Content relating to minority identity experience i.e. queer content, content about Black experiences
- [ ] Political content
- [ ] Content about violence that is intended to shock or disgust
- [ ] Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime
- [ ] Content relating to a social justice movement, for example feminism or anti-racism
- [ ] Curse words
- [ ] Hate speech
- [ ] Sex related content for a non-erotic purpose i.e. that is intended to educate
- [ ] _____ Other (please describe)

Please select how frequently your posts get removed when you post this kind of content:

	Never (this kind of content is never removed)	Sometimes	About half the time	Most of the time	Always (all my content about this topic is removed)
Content relating to a social justice movement, for example feminism or anti-racism	○	○	○	○	○
Hate speech	○	○	○	○	○
Content about violence that is intended to shock or disgust	○	○	○	○	○
Curse words	○	○	○	○	○
Sex related content for a non-erotic purpose i.e. that is intended to educate	○	○	○	○	○

	Never (this kind of content is never removed)	Sometimes	About half the time	Most of the time	Always (all my content about this topic is removed)
Content insulting or criticizing dominant group (e.g., men, white people)	○	○	○	○	○
Self-referential use of slur i.e. d*ke by a lesbian	○	○	○	○	○
Sex related content for an erotic purpose	○	○	○	○	○
Covid-related content	○	○	○	○	○
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime	○	○	○	○	○
Political content	○	○	○	○	○
Content relating to minority identity experience i.e. queer content, content about Black experiences	○	○	○	○	○
Content some may find offensive or inappropriate	○	○	○	○	○
Other _____	○	○	○	○	○

Are you paying attention?

- ○ No
- ○ Yes

Do you believe you have had content suppressed by TikTok?

- ○ No
- ○ Yes

Please select which of the types of content you believe have been suppressed by the platform.

I believe I have had posts suppressed that contained (you can select multiple):

- [ ] Covid-related content
- [ ] Content relating to minority identity experience i.e. queer content, content about Black experiences
- [ ] Content insulting or criticizing dominant group (e.g., men, white people)
- [ ] Sex related content for a non-erotic purpose i.e. that is intended to educate
- [ ] Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime
- [ ] Sex related content for an erotic purpose
- [ ] Curse words
- [ ] Self-referential use of slur i.e. d*ke by a lesbian
- [ ] Content about violence that is intended to shock or disgust
- [ ] Content relating to a social justice movement, for example feminism or anti-racism
- [ ] Political content
- [ ] Hate speech
- [ ] Content some may find offensive or inappropriate
- [ ] _____ Other (please describe)

Please select how frequently you believe your posts get suppressed when you post this kind of content:

	Never (this kind of content is never removed)	Sometimes	About half the time	Most of the time	Always (all my content about this topic is removed)
Content insulting or criticizing dominant group (e.g., men, white people)	○	○	○	○	○
Political content	○	○	○	○	○
Content some may find offensive or inappropriate	○	○	○	○	○
Sex related content for an erotic purpose	○	○	○	○	○
Covid-related content	○	○	○	○	○

	Never (this kind of content is never removed)	Sometimes	About half the time	Most of the time	Always (all my content about this topic is removed)
Sex related content for a non-erotic purpose i.e. that is intended to educate	○	○	○	○	○
Self-referential use of slur i.e. d*ke by a lesbian	○	○	○	○	○
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime	○	○	○	○	○
Content relating to minority identity experience i.e. queer content, content about Black experiences	○	○	○	○	○
Content relating to a social justice movement, for example feminism or anti-racism	○	○	○	○	○
Curse words	○	○	○	○	○
Content about violence that is intended to shock or disgust	○	○	○	○	○
Hate speech	○	○	○	○	○
Other ☐	○	○	○	○	○

We are interested in how algorithmic censorship make you feel. Please rate the extent to which you agree with the following statements. Algorithmic censorship makes me feel:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
glad, happy, joyful						☐
content, serene, peaceful						☐
ashamed, humiliated, disgraced						☐

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
scared, fearful, afraid						☐
angry, irritated, annoyed						☐
disgust, distaste, revulsion						☐
awe, wonder, amazement						☐
embarrassed, self-conscious, blushing						☐
amused, fun-loving, silly						☐
grateful, appreciative, thankful.						☐

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
love, closeness, trust						☐
repentant, guilty, blameworthy						☐
hate, distrust, suspicion						☐
stressed, nervous, overwhelmed						☐
contemptuous, scornful, disdainful						☐
hopeful, optimistic, encouraged						☐
interested, alert, curious						☐

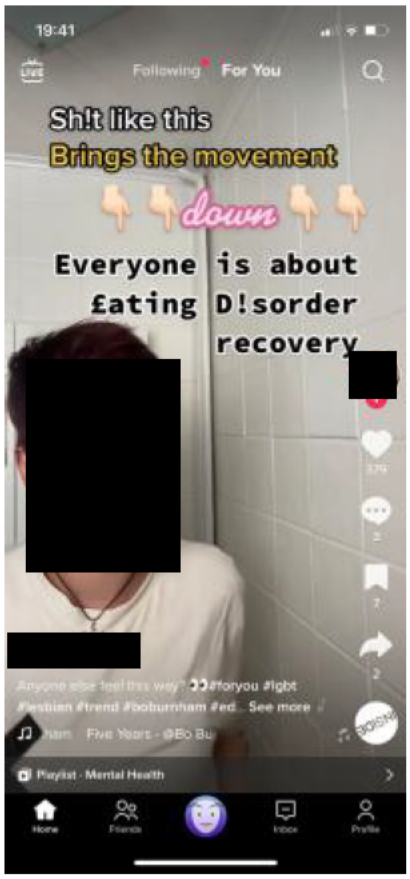
	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
sad, downhearted, unhappy						☐
inspired, uplifted, elevated						☐
proud, confident, self-assured						☐

Awareness of obfuscation

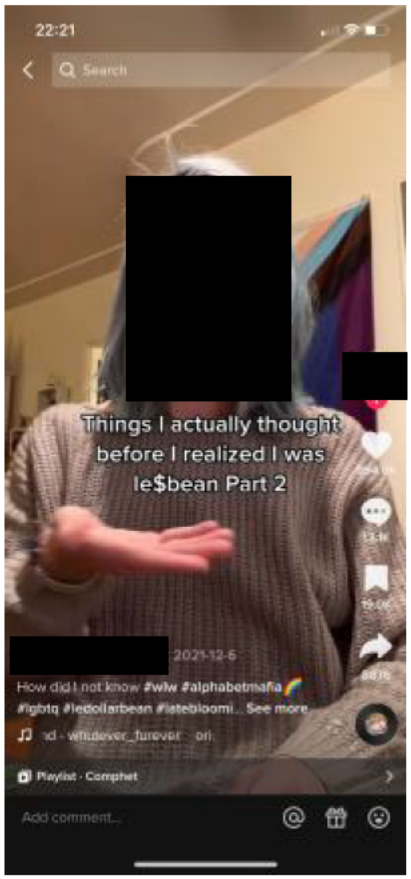
Some TikTok users are concerned about algorithmic censorship, which as we stated earlier refers to when an algorithm removes or suppresses the content they post. For example, users are concerned that the moderation system may censor content which includes words such as "lesbian" or "dead" in the text in the video or in the captions.

To avoid such censorship, some users employ different techniques to conceal what they are writing about. For example, users might write "le$bean" instead of "lesbian". This has been called "algospeak" and can be thought of as a kind of code, but we will use the term "obfuscation" which is used in research. People obfuscate certain words to conceal what they are saying to avoid algorithmic censorship.

Here are some examples of the techniques being used.

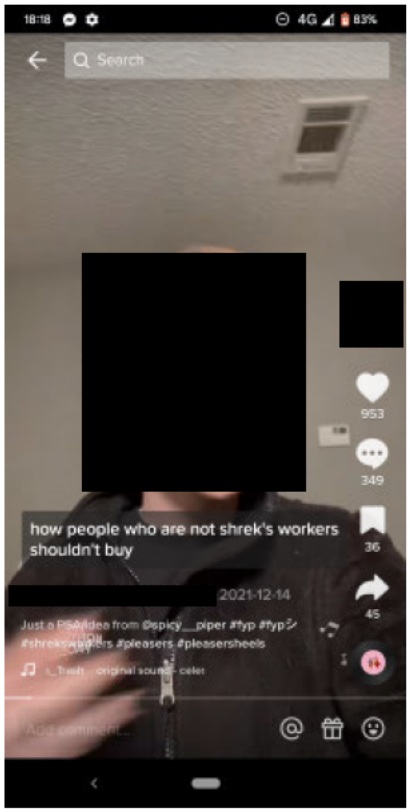


User substitutes numbers and symbols for letters.

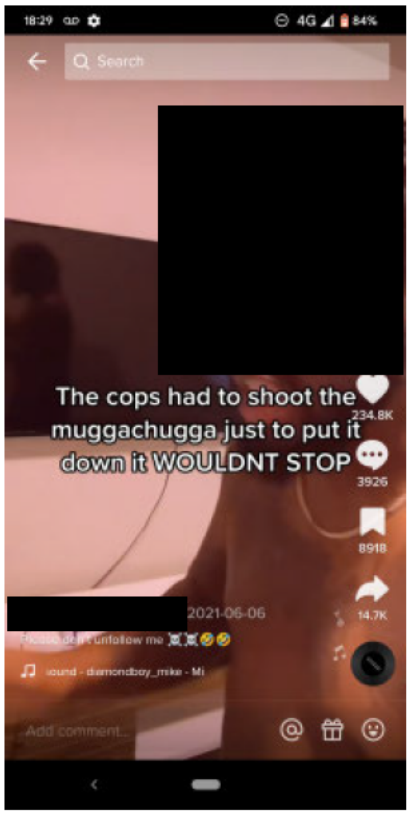


User substitutes a symbol for a letter. They also use the hashtag #alphabetmafia instead of #lgbtcommunity

User has changed the spelling of "white" to "yt".



User has substituted the word "sex" for "shrek's".



User has changed the word "motherfucker" with some sound changes and repeated letters.

Researchers have identified 7 key obfuscation techniques used by English language users (Calhoun & Fawcett, 2022). We will give specific examples of each technique below, taken from this research.

Users may also use these techniques for other reasons, such as making sure only certain TikTok users can understand what they are saying. We will ask about these motivations in the next section.

IMPORTANT: On mobile, please scroll right to see the explanations and examples which will help you understand these techniques.

via GIPHY

Non-technical Label	Explanation / description	Examples

Use of non-letters	Numbers, symbols, diacritics [marks added to letters like for *é*], emojis, spaces (including leaving the word "blank")	Gay > smoke crack > sm o ke cr a ck
Innovative subword combinations	Combining words and meaningful parts of words in new ways	dead > unalive [un + alive]
Word substitution (meaning or sound)	Replacing words with those with similar meanings or which sound similar	LGBTQ+ community > alphabet mafia faggot > baguette Sex work > shrek work
Sound changes	Spelling out a mispronounced version of the word	sexy > seggsy, sessy drugs > droogs faggot > fuhgoot
Spelling changes	Spelling out the word with homonyms Spelling out an initialism/acronym as words	circumcised > sircomesized LGBT > leg booty LGBT > los jibbities
Word substitution (structure)	Replacing word with one with the same number of syllables, with the emphasis put on matching syllables, and with some similar sounds	homophobic > hookedonphonics homophobia > cornucopia
Sound changes with repetition	Changing consonants to match	mother fucker > mugga chugga
Flip-flop version	Replacing word with phrase following the vowel pattern in "flip flop", "hip hop", "tip top" (or "zig zag", "tic tac" etc)	nipples > nip nops

Which of these techniques have you seen others using?

- ☐ Use of non-letters (gay >)
- ☐ Innovative subword combinations (unalive [un + alive])
- ☐ Sound changes with repetition (mother fucker > mugga chugga)
- ☐ Flip-flop version (nipples > nip nops)

- [ ] Word substitution (meaning or sound) (sex work > shrek work)
- [ ] Sound changes (sexy > seggsy)
- [ ] Spelling changes (LGBT > leg booty)
- [ ] Word substitution (structure) (homophobia > cornucopia)
- [ ] Other (please describe)
- [ ] None
- [ ] Unsure

How often do you understand what people's real intended meaning is when they use these techniques?

- ○ Never
- ○ Sometimes
- ○ About half the time
- ○ Most of the time
- ○ Always

Which of these techniques have you used yourself when you have posted content?

- [ ] Use of non-letters (gay >)
- [ ] Innovative subword combinations (unalive [un + alive])
- [ ] Word substitution (meaning or sound) (sex work > shrek work)
- [ ] Sound changes (sexy > seggsy)
- [ ] Spelling changes (LGBT > leg booty)
- [ ] Word substitution (structure) (homophobia > cornucopia)
- [ ] Sound changes with repetition (mother fucker > mugga chugga)
- [ ] Flip-flop version (nipples > nip nops)
- [ ] Other (please describe)
- [ ] None

How effective do you consider these techniques to be for avoiding algorithmic censorship?

	Not effective at all	Slightly effective	Moderately effective	Very effective	Extremely effective
	1	2	3	4	5
Use of non-letters (gay >)					
Innovative subword combinations (unalive [un + alive])					

	Not effective at all	Slightly effective	Moderately effective	Very effective	Extremely effective
	1	2	3	4	5
Word substitution (meaning or sound) (sex work > shrek work)					
Sound changes (sexy > seggsy)					
Spelling changes (LGBT > leg booty)					
Word substitution (structure) (homophobia > cornucopia)					
Sound changes with repetition (mother fucker > mugga chugga)					
Flip-flop version (nipples > nip nops)					

Is there anything you would like to add? If you have mentioned any additional techniques you have seen or used above, please indicate how effective you think they are for evading algorithmic censorship.

Feelings about obfuscation use

We are interested in how obfuscation techniques (when users conceal what they are writing about) make you feel. Please rate the extent to which you agree with the following statements.

Using or seeing obfuscation techniques makes me feel:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
awe, wonder, amazement					

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
content, serene, peaceful						☐
scared, fearful, afraid						☐
embarrassed, self-conscious, blushing						☐
angry, irritated, annoyed						☐
grateful, appreciative, thankful.						☐
amused, fun-loving, silly						☐
ashamed, humiliated, disgraced						☐
glad, happy, joyful						☐
disgust, distaste, revulsion						☐

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
hopeful, optimistic, encouraged						☐
inspired, uplifted, elevated						☐
stressed, nervous, overwhelmed						☐
interested, alert, curious						☐
proud, confident, self-assured						☐

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
love, closeness, trust					☐
contemptuous, scornful, disdainful					☐
hate, distrust, suspicion					☐
sad, downhearted, unhappy					☐
repentant, guilty, blameworthy					☐

Motivations (own use)

We are going to ask you about the types of content you post. Please select how frequently you post content about:

	Never (I never post content about this)	Sometimes	About half the time	Most of the time	Always (all my content is about this)
Sex related content for an erotic purpose	○	○	○	○	○
Hate speech	○	○	○	○	○
Content relating to minority identity experience i.e. queer content, content about Black experiences	○	○	○	○	○
Political content	○	○	○	○	○
Sex related content for a non-erotic purpose i.e. that is intended to educate	○	○	○	○	○
Content about violence that is intended to shock or disgust	○	○	○	○	○

	Never (I never post content about this)	Sometimes	About half the time	Most of the time	Always (all my content is about this)
Self-referential use of slur i.e. d*ke by a lesbian	○	○	○	○	○
Curse words	○	○	○	○	○
Covid-related content	○	○	○	○	○
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime	○	○	○	○	○
Content relating to a social justice movement, for example feminism or anti-racism	○	○	○	○	○
Content insulting or criticizing dominant group (e.g., men, white people)	○	○	○	○	○
Content some may find offensive or inappropriate	○	○	○	○	○

What other types of content do you post about? Please indicate how frequently you post this type of content.

We are going to ask you about your use of the obfuscation techniques we introduced you to earlier.
Please select how strongly you agree with each of the following statements relating to the types of content you are posting when you use obfuscation techniques.

I use obfuscation techniques when I post:

	Never	Sometimes	About half the time	Most of the time	Always
Political content	○	○	○	○	○
Content some may find offensive or inappropriate	○	○	○	○	○

	Never	Sometimes	About half the time	Most of the time	Always
Sex related content for a non-erotic purpose i.e. that is intended to educate	○	○	○	○	○
Sex related content for an erotic purpose	○	○	○	○	○
Covid-related content	○	○	○	○	○
Content insulting or criticizing dominant group (e.g., men, white people)	○	○	○	○	○
Content relating to a social justice movement, for example feminism or anti-racism	○	○	○	○	○
Content relating to minority identity experience i.e. queer content, content about Black experiences	○	○	○	○	○
Hate speech	○	○	○	○	○
Curse words	○	○	○	○	○
Self-referential use of slur i.e. d*ke by a lesbian	○	○	○	○	○
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime	○	○	○	○	○
Content about violence that is intended to shock or disgust	○	○	○	○	○

Are there any other types of content you post where you use obfuscation techniques? Please indicate how frequently you use obfuscation techniques for this kind of content

We are interested in your motivations for using obfuscation techniques when posting content. Please select how strongly you agree with the following statements.

I use obfuscation techniques:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
To feel part of the community of TikTok users					
So the algorithm does not censor my posts					
So human moderators can't understand what I am writing					
Because it's fun					
So only certain TikTok users can understand me					
To express my creativity					
To protest algorithmic moderation					
To show which communities I belong to					
Select somewhat agree					

Is there anything else you would like to tell us about when and why you use obfuscation techniques?

Beliefs about algorithmic censorship

Please select how strongly you agree with each of the following statements relating to your perception of algorithmic censorship by the platform. Please rate to what extent you agree with the following statements.

The **TikTok moderation algorithm censors** at least some posts about:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime					☐
Content relating to minority identity experience i.e. queer content, content about Black experiences					☐
Political content					☐
Hate speech					☐
Content about violence that is intended to shock or disgust					☐
Self-referential use of slur i.e. d*ke by a lesbian					☐
Sex related content for an erotic purpose					☐
Content relating to a social justice movement, for example feminism or anti-racism					☐
Sex related content for a non-erotic purpose i.e. that is intended to educate					☐
Content insulting or criticizing dominant group (e.g., men, white people)					☐
Covid-related content					☐

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
Curse words					
Content some may find offensive or inappropriate					

Is there anything that you would like to add about what you believe about the TikTok moderation algorithm?

Beliefs about community guidelines

Please select how strongly you agree with each of the following statements relating to your knowledge of the TikTok community guidelines.

TikTok community guidelines do not allow posts about:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
Content insulting or criticizing dominant group (e.g., men, white people)					
Content about violence that is not intended to shock or disgust i.e. reporting on a violent crime					
Sex related content for a non-erotic purpose i.e. that is intended to educate					
Hate speech					
Content relating to a social justice movement, for example feminism or anti-racism					

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
Sex related content for an erotic purpose					
Political content					
Content about violence that is intended to shock or disgust					
Content relating to minority identity experience i.e. queer content, content about Black experiences					
Self-referential use of slur i.e. d*ke by a lesbian					
Curse words					
Covid-related content					
Content some may find offensive or inappropriate					
Select somewhat disagree					

Is there anything that you would like to add about what you believe about the TikTok community guidelines?

We will now ask you some questions about your thoughts on TikTok's moderation process (both algorithmic and human moderation).

Sometimes users feel that content that seems to be in line with the community guidelines is censored. For example, the TikTok community guidelines do not disallow content about LGBT+ individuals (so long as this content does not violate their other policies such as the banning of sexually explicit content). However, LGBT+ creators talk about their content being suppressed or removed (Haimson et al., 2022; Lorenz

2022; Zeng and Kaye, 2022). We are interested in your beliefs about what is occurring when content that does not seem to violate the community guidelines is being removed.

When content is suppressed or removed that does not go against the community guidelines, this is because:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
	1	2	3	4	5
The algorithm has misunderstood the content / made a one-time mistake					
Other user(s) have reported the content					
TikTok has unpublished guidelines that are given to human moderators which are stricter					
TikTok has unpublished guidelines that are used to train the algorithm which are stricter					
Human moderators have their own opinions about what should be allowed on the platform					
The algorithm has not learned to follow the guidelines (it is not a good algorithm overall)					

Do you think there are any other reasons why this occurs (content that is in line with the community guidelines being removed by the algorithmic censorship system).

Demographic

What is your age? (Please give number of years)

What is your gender identity?

- ○ Male
- ○ Female
- ○ Non-binary
- ○ Other
- ○ Prefer not to say

Do you identify as trans?

- ○ Yes
- ○ No
- ○ Prefer not to say

What is your sexuality?

- ○ Straight / heterosexual
- ○ Gay / homosexual / lesbian
- ○ Bisexual / pansexual / bi+
- ○ Asexual / Acespec
- ○ Other
- ○ Prefer not to say

Do you consider yourself to be disabled?

- ○ No
- ○ Yes
- ○ Prefer not to say

What is your ethnic group?

- ○ Asian or Asian British

- ○ Black, African, Black British or Caribbean
- ○ Mixed or multiple ethnic groups
- ○ White
- ○ [ ] Another ethnic group (please specify)
- ○ Prefer not to say

What religious group do you belong to?

- ○ Buddhist
- ○ Christian
- ○ Hindu
- ○ Jewish
- ○ Muslim
- ○ Sikh
- ○ [ ] Other (please specify)
- ○ No religion
- ○ Prefer not to say

Political beliefs can be thought of on a spectrum from left to right. Where would you place yourself on this spectrum?

- ○ Left
- ○ Center left
- ○ Center
- ○ Center right
- ○ Right
- ○ I'm not sure
- ○ Prefer not to say

Think of this ladder as representing where people stand in the UK.

At the top of the ladder are the people who are the best off – those who have the most money, the most education, and the most respected jobs. At the bottom are the people who are the worst off – those who have the least money, least education, the least respected jobs, or no job. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to the people at the very bottom.

Where would you place yourself on this ladder?

Please indicate with a number (1-10) on which rung you think you stand at this time in your life relative to other people in the UK.

Do you identify as belonging to any other minority or marginalised group e.g. people with HIV, sex workers, care leavers, from a Gypsy, Roma & Traveller community etc

"Social identity" has been used to describe aspects of your identity related "to belonging to a certain social group, such as a certain race, gender or class" (Karizat et al., 2021).

We are interested in how connected to your social identity communities you feel on TikTok.

On TikTok...

	Strongly disgree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	
	1	2	3	4	5	
I feel like I am part of my communities						
I feel a connection with other members of my communities						
I feel accepted by other members of my community						
I feel respected by other members of my communities						
I feel valued by other members of my communities						

What is/are your first language(s)? (learned before age 5)

Bibliography

Abdalla, M. and Abdalla, M. (2021). The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 287–297, New York, NY, USA. Association for Computing Machinery.

Abercrombie, G., Vitsakis, N., Jiang, A., and Konstas, I. (2024). Revisiting Annotation of Online Gender-Based Violence. In Abercrombie, G., Basile, V., Bernadi, D., Dudy, S., Frenda, S., Havens, L., and Tonelli, S., editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 31–41, Torino, Italia. ELRA and ICCL.

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In Akiwowo, S., Vidgen, B., Prabhakaran, V., and Talat, Z., editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

An, H., Liu, X., and Zhang, D. (2022). Learning Bias-reduced Word Embeddings Using Dictionary Definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Anhorn, M. (2016). Two-Spirit. In *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pages 1–2. John Wiley & Sons.

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., and others (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv*. https://arxiv.org/abs/2404.09932 Accessed 16/07/2024.

Anzani, A., Lindley, L., Tognasso, G., Galupo, M. P., and Prunas, A. (2021). "Being Talked to Like I Was a Sex Toy, Like Being Transgender Was Simply for the Enjoyment of Someone Else": Fetishization and Sexualization of Transgender and Nonbinary Individuals. *Archives of Sexual Behavior*, 50(3):897–911.

Anzani, A., Siboni, L., Lindley, L., Paz Galupo, M., and Prunas, A. (2024). From Abstinence to Deviance: Sexual Stereotypes Associated With Transgender and Nonbinary Individuals. *Sexuality Research and Social Policy*, 21(1):27–43.

Are, C. (2020). How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist Media Studies*, 20(5):741–744.

Are, C. (2022). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22(8):2002–2019.

Are, C. (2023). The assemblages of flagging and de-platforming against marginalised content creators. *Convergence*, 30(2):922–937.

Bacchi, U. (2020). Tiktok apologises for censoring LGBT+ content. *Reuters*. https://www.reuters.com/article/britain-tech-lgbt-idUSL5N2GJ459 Accessed 05/08/2024.

Baker, P. (2019). *Fabulosa! The Story of Polari, Britain's Secret Gay Language*. Reaktion Books, Limited.

Bansal, H., Yin, D., Monajatipoor, M., and Chang, K.-W. (2022). How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barclay, D., Higgins, C., and Thompson, R. (1995). The partial least squares (PLS) approach to casual modeling: personal computer adoption ans use as an Illustration. *Technology Studies*, 2(2):285–309. Special Issue: Research Methodology.

Barikeri, S., Lauscher, A., Vulić, I., and Glavaš, G. (2021). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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), pages 1941–1955, Online. Association for Computational Linguistics.

Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. (2017). The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS)*, 2.

Battle, J. (2002). *Say it loud, I'm Black and I'm proud: Black pride survey 2000*. Policy Institute of the National Gay and Lesbian Task Force, New York.

Beer, D. (2022). The problem of researching a recursive society: Algorithms, data coils and the looping of the social. *Big Data & Society*, 9(2).

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Benjamin, G. (2021). What we do with data: a performative critique of data 'collection'. *Internet Policy Review*, 10(4).

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Bennett, S., Catanzariti, B., and Tollon, F. (2025). "Everybody knows what a pothole is": representations of work and intelligence in AI practice and governance. *AI & SOCIETY*.

Bestvater, S. (2024). How U.S. Adults Use TikTok. Technical report, Pew Research Center. https://www.pewresearch.org/internet/2024/02/22/how-u-s-adults-use-tiktok/ Accessed 13/03/2024.

Bhatt, S., Dev, S., Talukdar, P., Dave, S., and Prabhakaran, V. (2022). Recontextualizing Fairness in NLP: The Case of India. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1493–1504, New York, NY, USA. Association for Computing Machinery.

Biddle, S., Victor Ribeiro, Paulo, and Dias, T. (2020). TikTok told moderators: Suppress posts by the "ugly" and poor. *The Intercept*. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/ Accessed 2023-10-24.

Bieswanger, M. (2016). Aviation English: Two distinct specialised registers? In Schubert, C. and Sanchez-Stockhammer, C., editors, *Variational Text Linguistics: Revisiting Register in English*, pages 67–86. De Gruyter Mouton.

Billard, T. J. (2019). (No) Shame in the Game: The Influence of Pornography Viewing on Attitudes Toward Transgender People. *Communication research reports*, 36(1):45–56.

Bird, C., Ungless, E., and Kasirzadeh, A. (2023). Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 396–410, New York, NY, USA. Association for Computing Machinery.

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2022a). The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea. ACM.

Birhane, A. and Prabhu, V. (2021). Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, Los Alamitos, CA, USA. IEEE Computer Society.

Birhane, A., Prabhu, V. U., and Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv*. https://arxiv.org/abs/2110.01963 Accessed 19/01/2023.

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., and L. Dancy, C. (2022b). The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 948–958, New York, NY, USA. Association for Computing Machinery.

Blashki, K. and Nichol, S. (2005). Game Geek's Goss: Linguistic Creativity in Young Males within an Online University Forum (94//3 933k'5 9055ONEONE). *Australian Journal of Emerging Technologies and Society*, 3.

Bleiker, R., Campbell, D., Hutchison, E., and Nicholson, X. (2013). The visual dehumanisation of refugees. *Australian Journal of Political Science*, 48(4):398–416.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Blodgett, S. L. and O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv*. http://arxiv.org/abs/1707.00061 Accessed 08/10/2020.

Bodoff, D. and Ho, S. Y. (2016). Partial Least Squares Structural Equation Modeling Approach for Analyzing a Model with a Binary Indicator as an Endogenous Variable. *Communications of the Association for Information Systems*, 38:400–419.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. S. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh,

A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3663–3678. Curran Associates, Inc.

Borakati, A. (2021). Evaluation of an international medical E-learning course with natural language processing and machine learning. *BMC Medical Education*, 21(1).

Brack, M., Schramowski, P., Friedrich, F., Hintersdorf, D., and Kersting, K. (2023). The Stable Artist: Steering Semantics in Diffusion Latent Space. http://arxiv.org/abs/2212.06013 Accessed 23/08/2024.

Brandl, S., Cui, R., and Søgaard, A. (2022). How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Brewer, M. B. and Crano, W. D. (2014). Research Design and Issues of Validity. In Reis, H. T. and Judd, C. M., editors, *Handbook of Research Methods in Social and Personality Psychology*, pages 11–26. Cambridge University Press, 2nd edition.

Brown, A. (2021). TikTok influencer of color faced 'frustrating' obstacle trying to add the word 'black' to his creator Marketplace Bio. *Forbes*. https://www.forbes.com/sites/abrambrown/2021/07/07/tiktok-black-creators-creator-marketplace-black-lives-matter Accessed 2022-04-26.

Bucholtz, M. (1999). "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, 28(2):203–223.

Bucholtz, M. and Hall, K. (2004). Language and Identity. In Duranti, A., editor, *A Companion to Linguistic Anthropology*, pages 369–394. John Wiley & Sons, Incorporated, Williston, UK.

Bucholtz, M. and Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR. ISSN: 2640-3498.

Burchell, L. V. (2024). *Improving natural language processing for under-served languages through increased training data diversity*. PhD thesis, The University of Edinburgh.

Burke, P. J. and Stets, J. E. (2009). *Identity Theory*. Oxford University Press, Oxford, UK.

Cabello, L., Jørgensen, A. K., and Søgaard, A. (2023). On the Independence of Association Bias and Empirical Fairness in Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 370–378, Chicago IL USA. ACM.

Calabrese, A., Bevilacqua, M., Ross, B., Tripodi, R., and Navigli, R. (2021). AAA: Fair Evaluation for Abuse Detection Systems Wanted. In *13th ACM Web Science Conference 2021*, pages 243–252, Virtual Event United Kingdom. ACM.

Calder, J. (2019). From Sissy to Sickening: The Indexical Landscape of /s/ in SoMa, San Francisco. *Journal of Linguistic Anthropology*, 29(3):332–358.

Calhoun, K. and Fawcett, A. (2022). "They edited out her nip nops": Linguistic Innovation as Textual Censorship Avoidance on TikTok,". Presentation at the Annual Meeting of the Linguistic Society of America.

Calhoun, K. and Fawcett, A. (2023). "They edited out her nip nops": Linguistic innovation as textual censorship avoidance on TikTok. *Language@Internet*, 21(1).

Cao, Y., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., and Galstyan, A. (2022). On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Cao, Y. T. and Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Cercas Curry, A., Talat, Z., and Hovy, D. (2024). Impoverished Language Technology: The Lack of (Social) Class in NLP. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci,

A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8675–8682, Torino, Italia. ELRA and ICCL.

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., and De Choudhury, M. (2016). #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1201–1213, San Francisco California USA. ACM.

Cho, J., Zala, A., and Bansal, M. (2023). DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3020–3031, Paris, France. IEEE.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1):240:11324–240:11436.

Clark, O., Lee, M. M., Jingree, M. L., O'Dwyer, E., Yue, Y., Marrero, A., Tamez, M., Bhupathiraju, S. N., and Mattei, J. (2021). Weight Stigma and Social Media: Evidence and Public Health Solutions. *Frontiers in Nutrition*, 8.

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*, 34(4):739–766.

Collins, B. and Zadrozny, B. (2021). Anti-vaccine groups changing into 'dance parties' on Facebook to avoid detection. *NBC News*. https://www.nbcnews.com/tech/tech-

news/anti-vaccine-groups-changing-dance-parties-facebook-avoid-detection-rcna1480 Accessed 05/08/2024.

Cook, K. S. and Hegtvedt, K. A. (1983). Distributive Justice, Equity, and Equality. *Annual Review of Sociology*, 9(1):217–241.

Crawford, K. (2017). The Trouble with Bias. NeurIPS Keynote. https://www.youtube.com/watch?v=fMym_BKWQzk Accessed 31/01/2024.

Crawford, K. and Paglen, T. (2021). Excavating AI: the politics of images in machine learning training sets. *AI & SOCIETY*, 36(4):1105–1116.

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989:139–168.

Cresci, S., Trujillo, A., and Fagni, T. (2022). Personalized Interventions for Online Moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, pages 248–251, New York, NY, USA. Association for Computing Machinery.

Criado-Perez, C. (2019). *Invisible women: data bias in a world designed for men*. Abrams Press, New York. OCLC: 1048941266.

Croquet, P. (2024). A purple emoji instead of the word "rape": A paradox and symbol of speaking out on social media. https://www.lemonde.fr/en/pixels/article/2024/10/23/a-purple-emoji-instead-of-the-word-rape-a-paradox-and-symbol-of-speaking-out-on-social-media_6730250_13.html Accessed 06/10/2025.

Davis, J. L., Williams, A., and Yang, M. W. (2021). Algorithmic reparation. *Big Data & Society*, 8(2):20539517211044808. Publisher: SAGE Publications Ltd.

Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Lê Khac, P., Melas, L., and Ghosh, R. (2021). DALL·E Mini. [Computer software] https://github.com/borisdayma/dalle-mini Accessed 17/04/2024.

Delobelle, P., Tokpo, E. K., Calders, T., and Berendt, B. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of*

the Association for Computational Linguistics, pages 1693–1706. Association for Computational Linguistics.

Dennler, N., Ovalle, A., Singh, A., Soldaini, L., Subramonian, A., Tu, H., Agnew, W., Ghosh, A., Yee, K., Peradejordi, I. F., Talat, Z., Russo, M., and Pinhal, J. D. J. D. P. (2023). Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 375–386, New York, NY, USA. Association for Computing Machinery.

DePalma, R. and Atkinson, E. (2006). The sound of silence: Talking about sexual orientation and schooling. *Sexuality, Society and Learning*, 6(4):333–349.

Dev, S., Goyal, J., Tewari, D., Dave, S., and Prabhakaran, V. (2024). Building socio-culturally inclusive stereotype resources with community engagement. *Advances in Neural Information Processing Systems*, 36.

Dev, S., Jha, A., Goyal, J., Tewari, D., Dave, S., and Prabhakaran, V. (2023). Building Stereotype Repositories with Complementary Approaches for Scale and Depth. In Dev, S., Prabhakaran, V., Adelani, D., Hovy, D., and Benotti, L., editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., and Chang, K.-W. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devinney, H., Björklund, J., and Björklund, H. (2022). Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 862–872, New York, NY, USA. Association for Computing Machinery.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery. event-place: New Orleans, LA, USA.

Dolata, M., Feuerriegel, S., and Schwabe, G. (2021). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, pages 1–65.

DSIT (2025). Online Safety Act: explainer. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer Accessed 04/08/2025.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery.

Ehsan, U., Singh, R., Metcalf, J., and Riedl, M. (2022). The Algorithmic Imprint. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1305–1317, New York, NY, USA. Association for Computing Machinery.

Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Walsh, P., Groeneveld, D., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., and Dodge, J. (2024). What's In My Big Data? *arXiv*. http://arxiv.org/abs/2310.20707 Accessed 21/04/2024 Accepted at ICLR 2024 Spotlight.

Excell, E. and Al Moubayed, N. (2021). Towards Equal Gender Representation in the Annotations of Toxic Language Detection. In Costa-jussa, M., Gonen, H., Hardmeier, C., and Webster, K., editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 55–65, Online. Association for Computational Linguistics.

Feagin, J. (2013). *Systemic Racism: A Theory of Oppression*. Routledge, New York. Originally published 2006.

Feagin, J. and Bennefield, Z. (2014). Systemic racism and U.S. health care. *Social Science & Medicine*, 103:7–14.

Fischer, A. R. H., Tobi, H., and Ronteltap, A. (2011). When Natural met Social: A Review of Collaboration between the Natural and Social Sciences. *Interdisciplinary Science Reviews*, 36(4):341–358. Publisher: SAGE Publications.

Florini, S. (2014). Tweets, Tweeps, and Signifyin': Communication and Cultural Performance on "Black Twitter". *Television & New Media*, 15(3):223–237.

Foundation, H. R. C. (2022). An Epidemic of Violence 2022. Technical report, Human Rights Campaign Foundation. https://reports.hrc.org/an-epidemic-of-violence-2022 Accessed 26/05/2023.

Fredrickson, B. L. (2013). Chapter One - Positive Emotions Broaden and Build. In Devine, P. and Plant, A., editors, *Advances in Experimental Social Psychology*, volume 47, pages 1–53. Academic Press.

Fredrickson, B. L., Tugade, M. M., Waugh, C. E., and Larkin, G. R. (2003). What Good Are Positive Emotions in Crises? A Prospective Study of Resilience and Emotions Following the Terrorist Attacks on the United States on September 11th, 2001. *Journal of personality and social psychology*, 84(2):365–376.

Fredriksson, M. (2024). Understanding Bias in AI through Guardrails. *Choice 360*. https://www.choice360.org/libtech-insight/understanding-bias-in-ai-through-guardrails/ Accessed 01/08/2024.

Frick, N. R., Wilms, K. L., Brachten, F., Hetjens, T., Stieglitz, S., and Ross, B. (2021). The perceived surveillance of conversations through smart devices. *Electronic Commerce Research and Applications*, 47.

Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6):16–23.

Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In Doorn, N., Schuurbiers, D., van de Poel, I., and Gorman, M. E., editors, *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer Netherlands, Dordrecht.

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Frier, S. (2022). Why Facebook, Instagram look like TikTok. *Bloomberg*. https://www.bloomberg.com/news/articles/2022-07-27/why-facebook-instagram-look-like-tiktok Accessed 2024-07-17.

Fussell, S. (2019). How an Attempt at Correcting Bias in Tech Goes Wrong. *The Atlantic*. https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/ Accessed 28/03/2023.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, E., and Brewer, R. (2023). "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 205–216, New York, NY, USA. Association for Computing Machinery.

Galop (2021). Acephobia and anti-asexual hate crime. https://galop.org.uk/resource/acephobia-and-anti-asexual-hate-crime/ Accessed 14/08/2024.

Gani, J. K. and Khan, R. M. (2024). Positionality Statements as a Function of Coloniality: Interrogating Reflexive Methodologies. *International Studies Quarterly*, 68(2).

Gautam, V., Subramonian, A., Lauscher, A., and Keyes, O. (2024). Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP. In Faleńska, A., Basta, C., Costa-jussà, M., Goldfarb-Tarrant, S., and Nozza, D., editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for Datasets. *arXiv*. http://arxiv.org/abs/1803.09010 Accessed 31/03/2021.

Gehl, R. W., Moyer-Horner, L., and Yeo, S. K. (2017). Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science. *Television & New Media*, 18(6):529–547.

Ghaffary, S. (2021). How TikTok's hate speech detection tool set off a debate about racial bias on the app. *Vox*. https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media Accessed 2024-07-17.

Gibbs, M., Meese, J., Arnold, M., Nansen, B., and Carter, M. (2014). #Funeral and Instagram: death, social media, and platform vernacular. *Information, Communication & Society*, 18(3):255–268.

Glick, P. and Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., and Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. 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)*, pages 1926–1940, Online. Association for Computational Linguistics.

Goldfarb-Tarrant, S., Ungless, E., Balkir, E., and Blodgett, S. L. (2023). This prompt is measuring <mask>: evaluating bias evaluation in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Golec de Zavala, A., Waldzus, S., and Cypryanska, M. (2014). Prejudice towards gay men and a need for physical cleansing. *Journal of Experimental Social Psychology*, 54:1–10.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Good, C., Rattan, A., and Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, 102(4):700–717.

Grandhi, S. A., Plotnick, L., and Hiltz, S. R. (2019). Do I Stay or Do I Go?: Motivations and Decision Making in Social Media Non-use and Reversion. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–27.

Guo, W. and Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133, Virtual Event USA. ACM.

Guyan, K. (2021). Constructing a queer population? Asking about sexual orientation in Scotland's 2022 census. *Journal of Gender Studies*, 31(6):782–792.

Guyan, K. (2022). *Queer Data : Using Gender, Sex and Sexuality Data for Action*. Bloomsbury Studies in Digital Cultures. Bloomsbury Academic, London.

Haimson, O. L., Delmonaco, D., Nie, P., and Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2013). Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. *Long Range Planning*, 46(1-2):1–12.

Hair, J. F., Risher, J. J., Sarstedt, M., and Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1):2–24.

Hale, S. E. and Ojeda, T. (2018). Acceptable femininity? Gay male misogyny and the policing of queer femininities. *European Journal of Women's Studies*, 25(3):310–324.

Halliday, M. A. K. (1976). Anti-Languages. *American Anthropologist*, 78(3):570–584.

Harwell, D. (2022). How TikTok ate the internet. *Washington Post*. \url{https://www.washingtonpost.com/technology/interactive/2022/tiktok-popularity/} Accessed 2024-08-23.

Haslam, N. (2006). Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, 10(3):252–264.

Hataya, R., Bao, H., and Arai, H. (2023). Will Large-scale Generative Models Corrupt Future Datasets? In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20498–20508, Los Alamitos, CA, USA. IEEE Computer Society.

Henseler, J., Ringle, C. M., and Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In Sinkovics, R. R. and Ghauri, P. N., editors, *Advances in International Marketing*, volume 20, pages 277–319. Emerald Group Publishing Limited.

Hern, A. (2019a). Revealed: How Tiktok censors videos that do not please Beijing. *The Guardian*. https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing Accessed 11/03/2024.

Hern, A. (2019b). TikTok's local moderation guidelines ban pro-LGBT Content. *The Guardian*. https://www.theguardian.com/technology/2019/sep/26/tiktoks-local-moderation-guidelines-ban-pro-lgbt-content Accessed 11/03/2024.

Hille, J. J., Simmons, M. K., and Sanders, S. A. (2020). "Sex" and the Ace Spectrum: Definitions of Sex, Behavioral Histories, and Future Interest for Individuals Who Identify as Asexual, Graysexual, or Demisexual. *The Journal of Sex Research*, 57(7):813–823.

Hilte, L., Daelemans, W., and Vandekerckhove, R. (2018). Predicting Adolescents' Educational Track from Chat Messages on Dutch Social Media. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 328–334, Brussels, Belgium. Association for Computational Linguistics.

Hong, R., Agnew, W., Kohno, T., and Morgenstern, J. (2024). Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. http://arxiv.org/abs/2405.08209 Accessed 30/05/2024.

Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., and Szolovits, P. (2017). Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations. *IEEE Journal of Biomedical and Health Informatics*, 21(5):1449–1459. Conference Name: IEEE Journal of Biomedical and Health Informatics.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. (2020). Characterising Bias in Compressed Models. *arXiv*. http://arxiv.org/abs/2010.03058 Accessed 10/06/2024.

Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Hughes, A. (2022). Dall-e Mini: Creator explains blurred faces, going viral and the future of the project. *BBC Science Focus Magazine*. https://www.sciencefocus.com/news/dall-e-mini-creator-explains-blurred-faces-going-viral-and-the-future-of-the-project/ Accessed 17/07/2024.

Hughes, A., Trudgill, P., and Watt, D. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, Fifth Edition*. Routledge, London.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., and Prabhakaran, V. (2022). Evaluation Gaps in Machine Learning Practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1859–1876, New York, NY, USA. Association for Computing Machinery.

Hutto, C. J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8(1), pages 216–225.

HypeAuditor (2022). Distribution of TikTok audiences in the United Kingdom (UK) in 2021, by age and gender. Technical report, Statista. https://www.statista.com/statistics/1147635/distribution-of-tiktok-influencer-audience-by-age-and-gender-uk/ Accessed 17/07/2024.

Iantaffi, A. and Bockting, W. O. (2011). Views from both sides of the bridge? Gender, sexual legitimacy, and transgender people's experiences of relationships. *Culture, health & sexuality*, 13(3):355–370.

Ipsos (2023). Ipsos LGBT+ Pride 2023 Global Survey Report. Technical report, Ipsos. https://www.ipsos.com/en/pride-month-2023-9-of-adults-identify-as-lgbt Accessed 21/04/2024.

Jhaver, S., Zhang, A. Q., Chen, Q. Z., Natarajan, N., Wang, R., and Zhang, A. X. (2023). Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2):289:1–289:33.

Jones, R. H. (2023). Lip-synching and young people's everyday linguistic activism on TikTok. In *#YouthMediaLife & Friends*, pages 23–42. V&R unipress.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Incorporating Dialectal Variability for Socially Equitable Language Identification. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Kantola, J., Elomäki, A., Gaweda, B., Miller, C., Ahrens, P., and Berthet, V. (2022). "It's Like Shouting to a Brick Wall": Normative Whiteness and Racism in the European Parliament. *American Political Science Review*, pages 1–16.

Kantrowitz, A. (2023). The surprising consequences of every other app copying TikTok. *Slate*. https://slate.com/technology/2023/04/tiktok-facebook-instagram-youtube-reels-shorts-copying.html Accessed 05/08/2024.

Karizat, N., Delmonaco, D., Eslami, M., and Andalibi, N. (2021). Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):305:1–305:44.

Kaur, S., Kaul, P., and Zadeh, P. M. (2020). Study the Impact of COVID-19 on Twitter Users with respect to Social Isolation. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–6.

Keleg, A., Goldwater, S., and Magdy, W. (2023). ALDi: Quantifying the Arabic Level of Dialectness of Text. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Kelion, L. (2019). TikTok suppressed disabled users' videos. *BBC News*. https://www.bbc.co.uk/news/technology-50645345 Accessed 17/07/2024.

King, G., Pan, J., and Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 345.

Kiritchenko, S. and Mohammad, S. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Klassen, S. and Fiesler, C. (2022). "This Isn't Your Data, Friend": Black Twitter as a Case Study on Research Ethics for Public Data. *Social Media + Society*, 8(4).

Klug, D., Qin, Y., Evans, M., and Kaufman, G. (2021). Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. In *13th ACM Web Science Conference 2021*, pages 84–92, Virtual Event United Kingdom. ACM.

Knight, B. (2021). The banter, grief and joy of Black British twitter. *VICE*. https://www.vice.com/en/article/akgyka/the-banter-grief-and-joy-of-black-british-twitter Accessed 17/01/2024.

Knight, W. (2019). AI Is Biased. Here's How Scientists Are Trying to Fix It. *Wired*. https://www.wired.com/story/ai-biased-how-scientists-trying-fix/ Accessed: 26/03/2025.

Knight, W. (2022). Inside DALL-E Mini, the Internet's Favorite AI Meme Machine. *Wired*. https://www.wired.com/story/dalle-ai-meme-machine/ Accessed 28/02/2023.

Knowles, B., Fledderjohann, J., Richards, J. T., and Varshney, K. R. (2023). Trustworthy AI and the Logics of Intersectional Resistance. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 172–182, New York, NY, USA. Association for Computing Machinery.

Kolker, Z. M., Taylor, P. C., and Galupo, M. P. (2020). "As a Sort of Blanket Term": Qualitative Analysis of Queer Sexual Identity Marking. *Sexuality & Culture*, 24(5):1337–1357.

Köver, C. and Reuter, M. (2019). Discrimination: Tiktok Curbed Reach for people with disabilities. *Netzpolitik*. https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/ Accessed 17/07/2024.

LaCroix, T. and Luccioni, A. S. (2022). Metaethical Perspectives on 'Benchmarking' AI Ethics. *arXiv*. http://arxiv.org/abs/2204.05151 Accessed 23/08/2024.

Lane-Steele, L. (2011). Studs and Protest-Hypermasculinity: The Tomboyism within Black Lesbian Female Masculinity. *Journal of Lesbian Studies*, 15(4):480–492.

Larson, B. (2017). Gender as a Variable in Natural-Language Processing: Ethical Considerations. In Hovy, D., Spruit, S., Mitchell, M., Bender, E. M., Strube, M., and Wallach, H., editors, *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Lauscher, A., Crowley, A., and Hovy, D. (2022). Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lees, A., Borkan, D., Kivlichan, I., Nario, J., and Goyal, T. (2021). Capturing Covertly Toxic Speech via Crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 14–20, Online. Association for Computational Linguistics.

Lefkowitz, N. and Hedgcock, J. S. (2017). 13. Anti-language: Linguistic innovation, identity construction, and group affiliation among emerging speech communities. In Bell, N., editor, *Multiple Perspectives on Language Play*, pages 347–376. De Gruyter Mouton, Berlin, Boston.

Li, L., Ma, Z., and Cao, T. (2020). Leveraging social media data to study the community resilience of New York City to 2019 power outage. *International Journal of Disaster Risk Reduction*, 51:101776.

Li, Q., Wang, C., Liu, R., Wang, L., Zeng, D. D., and Leischow, S. J. (2018). Understanding Users' Vaping Experiences from Social Media: Initial Study Using Sentiment Opinion Summarization Techniques. *Journal of Medical Internet Research*, 20(8):e252.

Liang, C. (2021). Reflexivity, positionality, and disclosure in HCI. *Medium*. https://medium.com/@caliang/reflexivity-positionality-and-disclosure-in-hci-3d95007e9916 Accessed 23/11/2022.

Liang, C. A., Munson, S. A., and Kientz, J. A. (2021). Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People. *ACM Transactions on Computer-Human Interaction*, 28(2):1–47.

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., and Morency, L.-P. (2020). Towards Debiasing Sentence Representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Lidsky, D. (1998). Home on the Web. *PC Mag*, 17(15):116. https://books.google.com.mt/books?id=sdGz21L4yuAC Accessed 14/01/2022.

Lopez Torregrosa, L. (2021). Opinion | I'm Latina. Here's why I won't use the term Latinx. *NBC News*. https://www.nbcnews.com/think/opinion/many-latinos-say-latinx-offends-or-bothers-them-here-s-ncna1285916 Accessed 18/07/2024.

Lorenz, T. (2022). Internet 'Algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean'. *The Washington Post*. https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/ Accessed 26/04/2022.

Loria, S. (2018). textblob. (Version 0.15) [Computer software] https://textblob.readthedocs.io/en/dev/.

Lu, X., Lu, Z., and Liu, C. (2020). Exploring TikTok Use and Non-use Practices and Experiences in China. In Meiselwitz, G., editor, *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing*, Lecture Notes in Computer Science, pages 57–70. Springer International Publishing.

Luccioni, A. and Viviano, J. (2021). What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), pages 182–189, Online. Association for Computational Linguistics.

Lyu, Y., Cai, J., Callis, A., Cotter, K., and Carroll, J. M. (2024). "I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–15, New York, NY, USA. Association for Computing Machinery.

López, Q. (2022). Latinx, Latine, or Latino? 8 LGBTQ+ People Tell Us What They Prefer and Why. *Them*. https://www.them.us/story/latinx-latine-difference-definition Accessed 18/07/2024.

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., and Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1):52:1–52:26.

Mahelona, K., Leoni, G., Duncan, S., and Thompson, M. (2023). OpenAI's whisper is another case study in colonisation. *Papa Reo*. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/ Accessed 31/01/2024.

Marchiori Manerba, M., Stanczak, K., Guidotti, R., and Augenstein, I. (2024). Social Bias Probing: Fairness Benchmarking for Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14653–14671, Miami, Florida, USA. Association for Computational Linguistics.

Markl, N. (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.

Markl, N. and Lai, C. (2023). Everyone has an accent. In *INTERSPEECH 2023*, pages 4424–4427. ISCA.

Maronikolakis, A., Wisiorek, A., Nann, L., Jabbar, H., Udupa, S., and Schuetze, H. (2022). Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Meade, N., Poole-Dayan, E., and Reddy, S. (2022). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Meaney, J. A., Wilson, S. R., Chiruzzo, L., and Magdy, W. (2022). Don't Take It Personally: Analyzing Gender and Age Differences in Ratings of Online Humor. In Hopfgartner, F., Jaidka, K., Mayr, P., Jose, J., and Breitsohl, J., editors, *Social Informatics*, pages 20–33. Springer International Publishing.

Meng, N., Keküllüoğlu, D., and Vaniea, K. (2021). Owning and Sharing: Privacy Perceptions of Smart Speaker Users. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):45:1–45:29.

Mereish, E. H., Katz-Wise, S. L., and Woulfe, J. (2017). Bisexual-Specific Minority Stressors, Psychological Distress, and Suicidality in Bisexual Individuals: the Mediating Role of Loneliness. *Prevention Science*, 18(6).

Miceli, M., Posada, J., and Yang, T. (2022). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP):34:1–34:14.

Mina, A. X. (2014). Batman, Pandaman and the Blind Man: A Case Study in Social Change Memes and Internet Censorship in China. *Journal of Visual Culture*, 13(3):359–375.

Minkin, R. and Brown, A. (2021). Rising shares of U.S. adults know someone who is transgender or goes by gender-neutral pronouns. Technical report, Pew Research Center.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Mohammad, S. M. (2017). Challenges in Sentiment Analysis. In Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A., editors, *A Practical Guide to Sentiment Analysis*, Socio-Affective Computing, pages 61–83. Springer International Publishing.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Mousavi, R. and Gu, B. (2019). The Impact of Twitter Adoption on Lawmakers' Voting Orientations. *Information systems research*, 30(1):133–153.

Muradoglu, M., Horne, Z., Hammond, M. D., Leslie, S.-J., and Cimpian, A. (2021). Women—particularly underrepresented minority women—and early-career academics feel like impostors in fields that value brilliance. *Journal of Educational Psychology*, 114(5):1086. Publisher: US: American Psychological Association.

Nadal, K. L., Davidoff, K. C., and Fujii-Doe, W. (2014). Transgender Women and the Sex Work Industry: Roots in Systemic, Institutional, and Interpersonal Discrimination. *Journal of Trauma & Dissociation*, 15(2):169–183.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s):1–42.

Nichol, A. (2022). DALL·E 2 pre-training mitigations. https://openai.com/research/dall-e-2-pre-training-mitigations Accessed 21/04/2024.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, page 26.

Nozza, D., Bianchi, F., Lauscher, A., and Hovy, D. (2022). Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

O'Brien, T. (2012). Compounding injustice: The cascading effect of algorithmic bias in risk assessments. *Geo. JL & Mod. Critical Race Persp.*, 13:39.

Ohlheiser, A. (2020). TikTok has become the soul of the LGBTQ Internet. *Washington Post*. https://www.washingtonpost.com/technology/2020/01/28/tiktok-has-become-soul-lgbtq-internet/ Accessed 20/10/2023.

Ohlheiser, A. (2021). Welcome to TikTok's endless cycle of censorship and mistakes. *MIT Technology Review*. https://www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle/ Accessed 17/07/2024.

ONS (2022). Census 2021. https://www.ons.gov.uk/census Accessed 18/07/2024.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. (2021). Accumulative Poisoning Attacks on Real-time Data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2899–2912. Curran Associates, Inc.

Pasek, J. (2018). anesrake. (Version 0.80) [Computer software] https://cran.r-project.org/web/packages/anesrake/index.html.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E. L., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Phillips, T., Taylor, J., Narain, E., and Chandler, P. (2021). Selling Authentic Happiness: Indigenous wellbeing and romanticised inequality in tourism advertising. *Annals of Tourism Research*, 87:103115.

Poria, S., Hazarika, D., Majumder, N., and Mihalcea, R. (2020). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, pages 1–1. Conference Name: IEEE Transactions on Affective Computing.

Preston, C. C. and Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1–15.

Pérez-Pérez, M., Pérez-Rodríguez, G., Fdez-Riverola, F., and Lourenço, A. (2019). Using Twitter to Understand the Human Bowel Disease Community: Exploratory Analysis of Key Topics. *Journal of Medical Internet Research*, 21(8):e12610.

Qadri, R., Shelby, R., Bennett, C. L., and Denton, R. (2023). AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 506–517, New York, NY, USA. Association for Computing Machinery.

Qian, R., Ross, C., Fernandes, J., Smith, E. M., Kiela, D., and Williams, A. (2022). Perturbation Augmentation for Fairer NLP. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Razavi, S. Z. and Rahbari, M. (2020). Understanding Reactions to Natural Disasters: a Text Mining Approach to Analyze Social Media Content. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–7.

Renwick, T. and Barbosa, D. (2021). Detection and Identification of Obfuscated Obscene Language with Character Level Transformers. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Rice, D. R., Hudson, S.-k. T. J., and Noll, N. E. (2022). Gay = STIs? Exploring gay and lesbian sexual health stereotypes and their implications for prejudice and discrimination. *European Journal of Social Psychology*, 52(2):326–341. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2793.

Rigolot, C. (2020). Transdisciplinarity as a discipline and a way of being: complementarities and creative tensions. *Humanities and Social Sciences Communications*, 7(1):1–5.

Rimes, K. A., Goodship, N., Ussher, G., Baker, D., and West, E. (2019). Non-binary and binary transgender youth: Comparison of mental health, self-harm, suicidality, substance use and victimization experiences. *International Journal of Transgenderism*, 20(2-3):230–240.

Ringle, C. M., Wende, S., and Becker, J.-M. (2022). SmartPLS 4. [Computer software] https://www.smartpls.com.

Robertson, A., Magdy, W., and Goldwater, S. (2021). Black or White but Never Neutral: How Readers Perceive Identity from Yellow or Skin-toned Emoji. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–23.

Robila, M. and Robila, S. A. (2020). Applications of Artificial Intelligence Methodologies to Behavioral and Social Sciences. *Journal of Child and Family Studies*, 29(10):2954–2966.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Ross, A., Hughes, C., Ungless, E. L., and Lai, C. (2025). Conveying Gender Through Speech: Insights from Trans Men. In *Interspeech 2025*, pages 674–678.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan, F., Fritz, A., and Impiombato, D. (2020). TikTok and WeChat - Curating and controlling global information flows. Technical Report 37/2020, Australian Strategic Policy Institute. https://www.aspi.org.au/report/tiktok-wechat Accessed 17/01/2022.

Röttger, P., Vidgen, B., Nguyen, D., Talat, Z., Margetts, H., and Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 41–58, Online. Association for Computational Linguistics.

Saifee, D. H., Zheng, Z. E., Bardhan, I. R., and Lahiri, A. (2020). Are Online Reviews of Physicians Reliable Indicators of Clinical Outcomes? A Focus on Chronic Disease Management. *Information Systems Research*, 31(4):1282–1300.

Salinas Jr., C. (2020). The Complexity of the "x" in Latinx: How Latinx/a/o Students Relate to, Identify With, and Understand the Term Latinx. *Journal of Hispanic Higher Education*, 19(2):149–168.

Salminen, J., Jung, S.-g., Chowdhury, S., and Jansen, B. J. (2020). Analyzing Demographic Bias in Artificially Generated Facial Pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada. ACM.

Sanchez, A., Ross, A., and Markl, N. (2024). Beyond The Binary: Limitations and Possibilities of Gender-Related Speech Technology Research. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 526–532.

Sanders, E. B.-N. (2002). From user-centered to participatory design approaches. In *Design and the Social Sciences*. CRC Press.

Santy, S., Liang, J., Le Bras, R., Reinecke, K., and Sap, M. (2023). NLPositionality: Characterizing Design Biases of Datasets and Models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Schick, T., Udupa, S., and Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Schmahl, K. G., Viering, T. J., Makrodimitris, S., Naseri Jahfari, A., Tax, D., and Loog, M. (2020). Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103, Online. Association for Computational Linguistics.

Schnoebelen, T. (2017). The carrots and sticks of ethical NLP. *Medium*. https://medium.com/@TSchnoebelen/ethics-and-nlp-some-further-thoughts-53bd7cc3ff69 Accessed 26/03/2025.

Schwartz, L. (2022). Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Schwartz, R. and Stanovsky, G. (2022). On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Seaver, N. (2019). Knowing Algorithms. In Vertesi, J. and Ribes, D., editors, *DigitalSTS*, pages 412–422. Princeton University Press.

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Sen, I., Samory, M., Flöck, F., Wagner, C., and Augenstein, I. (2021). How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs? In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seshadri, P., Singh, S., and Elazar, Y. (2024). The Bias Amplification Paradox in Text-to-Image Generation. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6367–6384, Mexico City, Mexico. Association for Computational Linguistics.

Shah, D. S., Schwartz, H. A., and Hovy, D. (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.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Shen, H., DeVos, A., Eslami, M., and Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3405–3410, Hong Kong, China. Association for Computational Linguistics.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. http://arxiv.org/abs/2305.17493 Accessed 31/01/2024.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Sigurgeirsson, A. and Ungless, E. L. (2024). Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices. In *Interspeech 2024*, pages 3050–3054. ISCA.

Silva, A., Tambwekar, P., and Gombolay, M. (2021). Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Simmons, T. (2018). Gender isn't a Haircut: How representation of nonbinary people of color requires more than white androgyny. *Color Bloq*. This website is no longer available but content can be accessed using the Wayback machine.

Simpson, E. and Semaan, B. (2021). For You, or For "You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–34.

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2022). Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, Arlington VA USA. ACM.

Smart, A., Wang, D., Monk, E., Díaz, M., Kasirzadeh, A., Van Liemt, E., and Schmer-Galunder, S. (2024). Discipline and Label: A WEIRD Genealogy and Social Theory of Data Annotation. *arXiv*. arXiv:2402.06811 [cs].

Smith, E. M., Hall, M., Kambadur, M., Presani, E., and Williams, A. (2022). "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Social, W. A. (2024). Global Reddit user distribution by gender 2023 [Graph]. Technical report, We Are Social, & DataReportal, & Meltwater. https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/ Accessed 21/04/2024.

Soper, D. S. (2025). A-priori Sample Size Calculator for Structural Equation Models. Available from https://www.danielsoper.com/statcalc.

Starke, C., Baleis, J., Keller, B., and Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2).

Steed, R., Panda, S., Kobren, A., and Wick, M. (2022). Upstream Mitigation Is *N*ot All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Stewart, I., Chancellor, S., De Choudhury, M., and Eisenstein, J. (2017). #Anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4353–4361.

Stone, A. L. (2009). More than Adding a T: American Lesbian and Gay Activists' Attitudes towards Transgender Inclusion. *Sexualities*, 12(3):334–354.

Strengers, Y., Qu, L., Xu, Q., and Knibbe, J. (2020). Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA. ACM.

Struppek, L., Hintersdorf, D., Friedrich, F., Br, M., Schramowski, P., and Kersting, K. (2023). Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068.

Suresh, H. and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Tait, A. (2022). Are Tiktok algorithms changing how people talk about suicide? *Wired*. https://www.wired.co.uk/article/algorithms-suicide-unalive Accessed 05/08/2024.

Tal, Y., Magar, I., and Schwartz, R. (2022). Fewer Errors, but More Stereotypes? The Effect of Model Size on Gender Bias. In Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H., editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Talat, Z., Lulz, S., Bingel, J., and Augenstein, I. (2021). Disembodied Machine Learning: On the Illusion of Objectivity in NLP. *ArXiv*. http://arxiv.org/abs/2101.11974 Accessed 08/04/2021.

Tan, S., Joty, S., Varshney, L., and Kan, M.-Y. (2020). Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663. Conference Name: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) Place: Online Publisher: Association for Computational Linguistics.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558.

Thompson, A. (2017). Google's Sentiment Analyzer Thinks Being Gay Is Bad. *VICE*. https://www.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias Accessed 14/01/2020.

Thylstrup, N. and Talat, Z. (2020). Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *Social Science Research Network*. https://papers.ssrn.com/abstract=3709719 Accessed 10/06/2024.

Thylstrup, N. B., Hansen, K. B., Flyverbom, M., and Amoore, L. (2022). Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, 9(2):20539517221139785.

Tobi, H. and Kampen, J. K. (2018). Research design: the methodology for interdisciplinary research framework. *Quality & Quantity*, 52(3):1209–1225.

Tulshyan, R. and Burey, J.-A. (2021). Stop telling women they have imposter syndrome. *Harvard Business Review*. https://hbr.org/2021/02/stop-telling-women-they-have-imposter-syndrome Accessed 01/08/2024.

Tway, P. (1975). Workplace isoglosses: Lexical variation and change in a factory setting. *Language in Society*, 4(2):171–183.

Táíwò, O. (2021). Being-in-the-Room Privilege: Elite Capture and Epistemic Deference. *The Philosopher*, 108(4). https://www.thephilosopher1923.org/post/being-in-the-room-privilege-elite-capture-and-epistemic-deference Accessed 26/03/2025.

Ungless, E., Ross, B., and Belle, V. (2023a). Potential Pitfalls with Automatic Sentiment Analysis: The Example of Queerphobic Bias. *Social science computer review*, 41(6):2211–2229.

Ungless, E., Ross, B., and Lauscher, A. (2023b). Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Ungless, E. L., Dev, S., Bennett, C. L., Gulotta, R., Bastings, J., and Denton, R. (2025a). Amplifying Trans and Nonbinary Voices: A Community-Centred Harm Taxonomy for LLMs. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20503–20535, Vienna, Austria. Association for Computational Linguistics.

Ungless, E. L., Markl, N., and Ross, B. (2025b). Experiences of Censorship on TikTok Across Marginalised Identities. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1952–1965.

Ungless, E. L., Markl, N., and Ross, B. (2025c). Le$bean or lesbian? A survey of marginalised users' motivations for obfuscation on TikTok. *Behaviour & Information Technology*, pages 1–26.

Ungless, E. L., Rafferty, A., Nag, H., and Ross, B. (2022). A Robust Bias Mitigation Procedure Based on the Stereotype Content Model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science*

(NLP+CSS), pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Ungless, E. L., Vitsakis, N., Talat, Z., Garforth, J., Ross, B., Onken, A., Kasirzadeh, A., and Birch, A. (2024). Ethics Whitepaper: Whitepaper on Ethical Research into Large Language Models. http://arxiv.org/abs/2410.19812 Accessed 08/12/2024.

Ungless, E. L., Vitsakis, N., Talat, Z., Garforth, J., Ross, B., Onken, A., Kasirzadeh, A., and Birch, A. (2025d). The Only Way is Ethics: A Guide to Ethical Research with Large Language Models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8992–9005, Abu Dhabi, UAE. Association for Computational Linguistics.

Utama, P. A., Moosavi, N. S., and Gurevych, I. (2020). Towards Debiasing NLU Models from Unknown Biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Valentine, V. (2016). Non-binary people's experiences in the UK. Technical report, The Equality Network: Scottish Trans. http://www.scottishtrans.org/non-binary Accessed 24/11/2022.

Vaterlaus, J. M. and Winter, M. (2021). TikTok: an exploratory study of young adults' uses and gratifications. *The Social Science Journal (Fort Collins)*, pages 1–20.

Velkova, J. and Kaun, A. (2021). Algorithmic resistance: media practices and the politics of repair. *Information, Communication & Society*, 24(4):523–540.

Venté, B. (2023). "Foundation Models" Are Not Foundational and Never Have Been. *AI Mind*. https://pub.aimind.so/foundation-models-f9c0916497ac Accessed 13/03/2025.

Vynck, G. D. and Tiku, N. (2024). Google takes down Gemini AI image generator. Here's what you need to know. *Washington Post*. https://www.washingtonpost.com/technology/2024/02/22/google-gemini-ai-image-generation-pause/ Accessed 19/08/2024.

Wallach, H. (2014). Big Data, Machine Learning, and the Social Sciences. *Medium*. https://hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d Accessed 26/03/2025.

Walter, M., Lovett, R., Maher, B., Williamson, B., Prehn, J., Bodkin-Andrews, G., and Lee, V. (2021). Indigenous Data Sovereignty in the Era of Big Data and Open Data. *Australian Journal of Social Issues*, 56(2):143–156.

Wang, C., Wang, K., Bian, A., Islam, R., Keya, K. N., Foulds, J., and Pan, S. (2022). Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. In *27th International Conference on Intelligent User Interfaces*, pages 134–147, Helsinki Finland. ACM.

Wang, R., Harper, F. M., and Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA. ACM.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., and Petrov, S. (2021). Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv*. http://arxiv.org/abs/2010.06032.

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

West, S. M., Whittaker, M., and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. Technical report, AI Now Institute.

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, 9(6):476–487.

Widder, D. G. (2024). Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1295–1304, New York, NY, USA. Association for Computing Machinery.

Widder, D. G., West, S., and Whittaker, M. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. *Social Science Re-*

search Network. https://ssrn.com/abstract=4543807 Accessed 02/02/2024 Accepted to appear in Nature.

Wolf, E. J., Harrington, K. M., Clark, S. L., and Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, 73(6):913–934. Publisher: SAGE Publications Inc.

Wolfe, R., Yang, Y., Howe, B., and Caliskan, A. (2023). Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1174–1185, New York, NY, USA. Association for Computing Machinery.

Woods, M. (2021). It's queers versus TikTok moderation. *Xtra*. https://xtramagazine.com/power/tiktok-censorship-queer-moderation-200629 Accessed 04/10/2022.

Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., and Klein, D. (2021). Detoxifying Language Models Risks Marginalizing Minority Voices. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Yeo, C. and Chen, A. (2020). Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 107–109, Seattle, United States. Association for Computational Linguistics.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y. (2022). Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. http://arxiv.org/abs/2206.10789 Accessed 08/02/2023.

Yu, Z., Sen, I., Assenmacher, D., Samory, M., Fröhling, L., Dahn, C., Nozza, D., and Wagner, C. (2024). The Unseen Targets of Hate: A Systematic Review of Hateful Communication Datasets. *Social Science Computer Review*, 0(0).

Zane, Z. (2019). Should Straight People Attend LGBTQ Pride? *Rolling Stone*. https://www.rollingstone.com/culture/culture-features/should-straight-people-attend-lgbtq-pride-666076/ Accessed 14/08/2024.

Zeng, J. and Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, 14(1):79–95.

Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., and Zhao, T. (2020). Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Zhang, S., Zhang, X., Zhang, W., and Søgaard, A. (2021). Sociolectal Analysis of Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhao, D., Wang, A., and Russakovsky, O. (2021). Understanding and Evaluating Racial Biases in Image Captioning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14810–14820, Los Alamitos, CA, USA. IEEE Computer Society.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. (2024). WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*.

Zhong, C.-B., Strejcek, B., and Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology*, 46(5):859–862.

Zhou, Z., Yang, M., and Jin, X.-L. (2018). Differences in the Reasons of Intermittent versus Permanent Discontinuance in Social Media: An Exploratory Study in Weibo. *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 493–502.

Zimman, L. and Hayworth, W. (2020). How we got here: Short-scale change in identity labels for trans, cis, and non-binary people in the 2000s. *Proceedings of the Linguistic Society of America*, 5(1):499–513.

Łozowski, P. (2017). Dictionaries and culture. In *The Routledge Handbook of Lexicography*. Routledge.