

# Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities

Nenad Tomasev  
nenadt@deepmind.com  
DeepMind  
London, United Kingdom

Jackie Kay\*  
kayj@deepmind.com  
DeepMind  
London, United Kingdom

Kevin R. McKee  
kevinrmckee@deepmind.com  
DeepMind  
London, United Kingdom

Shakir Mohamed  
shakir@deepmind.com  
DeepMind  
London, United Kingdom

## ABSTRACT

Advances in algorithmic fairness have largely omitted sexual orientation and gender identity. We explore queer concerns in privacy, censorship, language, online safety, health and employment to study the positive and negative effects of artificial intelligence on queer communities. These issues underscore the need for new directions in fairness research that take into account a multiplicity of considerations, from privacy preservation, context sensitivity and process fairness, to an awareness of sociotechnical impact and the increasingly important role of inclusive and participatory research processes. Most current approaches for algorithmic fairness assume that the target characteristics for fairness—frequently, race and legal gender—can be observed or recorded. Sexual orientation and gender identity are prototypical instances of *unobserved* characteristics, which are frequently missing, unknown or fundamentally unmeasurable. This paper highlights the importance of developing new approaches for algorithmic fairness that break away from the prevailing assumption of observed characteristics.

## CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Social and professional topics** → **Sexual orientation**; **Gender**; **Computing / technology policy**.

## KEYWORDS

algorithmic fairness, queer communities, sexual orientation, gender identity, machine learning, marginalised groups

### ACM Reference Format:

Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, May 19–21, 2021, Virtual Event, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3461702.3462540>

\*Also with Centre for Artificial Intelligence, University College London.



This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES '21, May 19–21, 2021, Virtual Event, USA.  
© 2021 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-8473-5/21/05.  
<https://doi.org/10.1145/3461702.3462540>

## 1 INTRODUCTION

As the field of algorithmic fairness has matured, the ways in which machine learning researchers and developers operationalise approaches for fairness have expanded in scope and applicability. Fairness researchers have made important advances and demonstrated how the risks of algorithmic systems are imbalanced across different characteristics of the people who are analysed and affected by classifiers and decision-making systems [15, 57]. Progress has been particularly strong with respect to race and legal gender.<sup>1</sup> Fairness studies have helped to draw attention to racial bias in recidivism prediction [9], expose racial and gender bias in facial recognition [32], reduce gender bias in language processing [26, 124], and increase the accuracy and equity of decision making for child protective services [40].

Algorithms have moral consequences for queer communities, too. However, algorithmic fairness for queer individuals and communities remains critically underexplored. In part, this stems from the unique challenges posed by studying sexual orientation and gender identity. Most definitions of algorithmic fairness share a basis in norms of egalitarianism [15, 20].<sup>2</sup> For example, classification parity approaches to fairness aim to equalise predictive performance measures across groups, whereas anti-classification parity approaches rely on the omission of protected attributes from the decision making process to ensure different groups receive equivalent treatment [43]. An inherent assumption of these approaches is that the protected characteristics are known and available within datasets. Sexual orientation and gender identity are prototypical examples of *unobserved* characteristics, presenting challenging obstacles for fairness research [8, 78].

This paper explores the need for *queer fairness* by reviewing the experiences of technological impacts on queer communities. For our discussion, we define ‘queer’ as ‘possessing non-normative sexual identity, gender identity, and/or sexual characteristics’. We consider

<sup>1</sup>Throughout this paper, we distinguish between ‘legal gender’ (the gender recorded on an individual’s legal documents, often assigned to them at birth by the government, physicians or their parents) and ‘gender identity’ (an individual’s personal feelings and convictions about their gender; [30]).

<sup>2</sup>It is worth noting that certain algorithmic domains supplement egalitarian concerns with additional ethical values and principles. For example, fairness assessments of healthcare applications typically incorporate beneficence and non-maleficence, two principles central to medical ethics [16].

this to include lesbian, gay, bisexual, pansexual, transgender, and asexual identities—among others.<sup>3</sup>

The focus on queer communities is important for several reasons. Given the historical oppression and contemporary challenges faced by queer communities, there is a substantial risk that artificial intelligence (AI) systems will be designed and deployed unfairly for queer individuals. Compounding this risk, sensitive information for queer people is usually not available to those developing AI systems, rendering the resulting unfairness unmeasurable from the perspective of standard group fairness metrics. Despite these issues, fairness research with respect to queer communities is an understudied area. Ultimately, the experiences of queer communities can reveal insights for algorithmic fairness that are transferable to a broader range of characteristics, including disability, class, religion, and race.

This paper aims to connect ongoing efforts to strengthen queer communities in AI research [3, 4, 83] and sociotechnical decision making [77, 98, 99, 122] with recent advances in fairness research, including promising approaches to protecting unobserved characteristics. This work additionally advocates for the expanded inclusion of queer voices in fairness and ethics research, as well as the broader development of AI systems. We make three contributions in this paper:

- (1) Expand on the promise of AI in empowering queer communities and supporting LGBTQ+ rights and freedoms.
- (2) Emphasise the potential harms and unique challenges raised by the sensitive and unmeasurable aspects of identity data for queer people.
- (3) Based on use cases from the queer experience, establish requirements for algorithmic fairness on unobserved characteristics.

## 2 CONSIDERATIONS FOR QUEER FAIRNESS

To emphasise the need for in-depth study of the impact of AI on queer communities around the world, we explore several interconnected case studies of how AI systems interact with sexual orientation and gender identity. These case studies highlight both potential benefits and risks of AI applications for queer communities. In reviewing these cases, we hope to motivate the development of technological solutions that are inclusive and beneficial to everyone. Importantly, these case studies will demonstrate cross-cutting challenges and concerns raised by unobserved and missing characteristics, such as preserving privacy, supporting feature imputation, context-sensitivity, exposing coded inequity, participatory engagement, and sequential and fair processes.

### 2.1 Privacy

Sexual orientation and gender identity are highly private aspects of personal identity. Outing queer individuals—by sharing or exposing their sexual orientation or gender identity without their prior consent—can not only lead to emotional distress, but also risk serious physical and social harms, especially in regions where

queerness is openly discriminated against [161], criminalised [49] or persecuted [136]. Privacy violations can thus have major consequences for queer individuals, including infringement upon their basic human rights [7, 28], denial of employment and education opportunities, ill-treatment, torture, sexual assault, rape, and extrajudicial killings.

**2.1.1 Promise.** Advances in privacy-preserving machine learning [27, 81, 118] present the possibility that the queer community might benefit from AI systems while minimising the risk of information leakage. Researchers have proposed adversarial filters [103, 169] to obfuscate sensitive information in images and speech shared online while reducing the risks of re-identification.

Still, challenges remain [144]. More research is needed to ensure the robustness of the adversarial approaches. The knowledge gap between the privacy and machine-learning research communities must be bridged for these approaches to achieve the desired effects. This will ensure that the appropriate types of protections are included in the ongoing development of AI solutions [5].

**2.1.2 Risks.** A multitude of privacy risks arise for queer people from the applications of AI systems. We focus in particular on the categorisation of identity from sensitive data, the ethical risk of surveillance, and invasions of queer spaces.

In 2017, Stanford researchers attempted to build an AI ‘gaydar’, a computer vision model capable of guessing a person’s sexual orientation from images [162]. The resulting algorithm, a logistic regression model trained on 35,326 facial images, achieved a high reported accuracy in identifying self-reported sexual orientation across both sexes. The results of this study have since been questioned, largely on the basis of a number of methodological and conceptual flaws that discredit the performance of the system [59]. Other algorithms designed to predict sexual orientation have suffered similar methodological and conceptual deficiencies. A recently released app claimed to be able to quantify the evidence of one’s non-heterosexual orientation based on genetic data, for example [18], largely obfuscating the limited ability of genetic information to predict sexual orientation [58].

Though these specific efforts have been flawed, it is plausible that in the near future algorithms could achieve high accuracy, depending on the data sources involved. Behavioural data recorded online present particular risks to the privacy of sexual orientation and gender identity: after all, the more time people spend online, the greater their digital footprint. AI ‘gaydars’ relying on an individual’s recorded interests and interactions could pose a serious danger to the privacy of queer people. In fact, as a result of the long-running perception of the queer community as a profitable ‘consumer group’ by business and advertisers alike, prior efforts have used online data to map ‘queer interests’ in order to boost sales and increase profits [140]. In at least one instance, researchers have attempted to use basic social media information to reconstruct the sexual orientation of users [19].

The ethical implications of developing such systems for queer communities are far-reaching, with the potential of causing serious harms to affected individuals. Prediction algorithms could be deployed at scale by malicious actors, particularly in nations where homosexuality and gender nonconformity are punishable

<sup>3</sup>Throughout this paper, we use ‘queer’ and ‘LGBTQ+’ interchangeably. The heterogeneity of queer communities—and the complexity of the issues they face—preclude this work from being an exhaustive review of queer identity. As a result, there are likely perspectives that were not included in this manuscript, but that have an important place in broader discussions of queer fairness.

offences. In fact, in many such nations, authorities already use technology to entrap or locate queer individuals through social media and LGBTQ+ dating apps (e.g., [48]). Systems predicting sexual orientation may also exacerbate the pre-existing privacy risks of participating in queer digital spaces. There have been recorded cases of coordinated campaigns for outing queer people, resulting in lives being ruined, or lost due to suicide [54]. These malicious outing campaigns have until now been executed at smaller scales. However, recent developments in AI greatly amplify the potential scale of such incidents, endangering larger communities of queer people in certain parts of the world. Facial recognition technology [159] could be employed by malicious actors to rapidly identify individuals sharing their pictures online, whether publicly or in direct messages. Facial recognition could similarly be used to automatically identify people in captured recordings of protests, in queer nightclubs or community spaces, and other in-person social events. These possibilities highlight the potential dangers of AI for state-deployed surveillance technology. Chatbots have similarly been deployed to elicit private information on dating apps, compromising users' device integrity and privacy [109]. Existing bots are scripted, and therefore can usually be distinguished from human users after longer exchanges. Nonetheless, strong language models [31] threaten to exacerbate such privacy risks, given their ability to quickly adjust to the style of communication based on a limited number of examples. These language models amplify existing concerns around the collection of private information and the compromising of safe online spaces.

In addition to the direct risks to privacy, algorithms intended to predict sexual orientation and gender identity also perpetuate concerning ideas and beliefs about queerness. Systems using genetic information as the primary input, for example, threaten to reinforce biological essentialist views of sexual orientation and echo tenets of eugenics—a historical framework that leveraged science and technology to justify individual and structural violence against people perceived as inferior [121, 164]. More broadly, the design of predictive algorithms can lead to erroneous beliefs that biology, appearance or behaviour are the essential features of sexual orientation and gender identity, rather than imperfectly correlated causes, effects or covariates of queerness.

In sum, sexual orientation and gender identity are associated with key privacy concerns. Non-consensual outing and attempts to infer protected characteristics from other data thus pose ethical issues and risks to physical safety. In order to ensure queer algorithmic fairness, it will be important to develop methods that can improve fairness for marginalised groups without having direct access to group membership information. This could be achieved either through proxies [63] when there is sufficient contextual information in the data, or by implementing more general principles to ensure that similar individuals are treated similarly [51], under any plausible unobserved grouping [95].

## 2.2 Censorship

Multiple groups and institutions around the world impose unjust restrictions on the freedom of expression and speech of queer communities. This censorship is often justified by its supporters as 'preserving decency' and 'protecting the youth', but in reality leads

to the erasure of queer identity. Laws against 'materials promoting homosexuality' were established in the late 1980s in the United Kingdom and repealed as late as 2003 [34]. Nations that are considered major world powers have laws banning the portrayal of same-sex romances in television shows (e.g., China; [105]), the mention of homosexuality or transgender identities in public education (e.g., state-level laws in the United States; [72]), or any distribution of LGBT-related material to minors (e.g., Russia; [91]). Not only do such laws isolate queer people from their communities—particularly queer youth—they shame queerness as indecent behaviour, setting a precedent for further marginalisation and undermining of human rights. Many queer content producers in such nations have argued that their online content is being restricted and removed at the detriment of queer expression and sex positivity, as well as at the cost of their income [167].

**2.2.1 Promise.** AI systems may be effectively used to mitigate censorship of queer content. Machine learning has been used to analyse and reverse-engineer patterns of censorship. A study of 20 million tweets from Turkey employed machine learning to show that the vast majority of censored tweets contained political content [150]. A statistical analysis of Weibo posts and Chinese-language tweets uncovered a set of charged political keywords present in posts with anomalously high deletion rates [14]. Further study of censorship could be key to drawing the international community's attention to human rights violations. It might also be useful for empowering affected individuals to circumvent these unfair restrictions. For example, a comparative analysis of censorship across platforms might help marginalised communities identify safe spaces where freedom of expression is least obstructed, as well as provide evidence and help coordinate action against platforms responsible for discriminatory censorship. Nonetheless, no large-scale study of censored queer content has yet been conducted, rendering these forward-looking technical applications somewhat speculative at the moment.

**2.2.2 Risk.** Although we believe machine learning can be used to combat censorship, tools for detecting queer digital content can be abused to enforce censorship laws or heteronormative cultural attitudes. As social network sites, search engines, and other media platforms adopt algorithms to moderate content at scale, the risk for unfair or biased censorship of queer content increases, and governing entities are empowered to erase queer identities from the digital sphere [41]. Automated content moderation systems are at risk of censoring queer expression even when the intention is benign, such as protecting users from verbal abuse. To help combat censorship restrictions and design fair content moderation systems, ML fairness researchers could investigate how to detect and analyse anomalous omission of information related to queer identity (or other protected characteristics) in natural language and video data.

Censorship often goes hand-in-hand with the distortion of facts. Recent advances in generative models have made the fabrication of digital content trivial, given enough data and computational power [38]. Malicious and dehumanising misinformation about the queer community has been used as justification for abuse and suppression throughout history, tracing back to medieval interpretations of ancient religious texts [52]. Technological and political solutions to the threat of misinformation are important for protecting

queer expression—as well as global democracy. The AI community has begun to develop methods to verify authentic data through, for example, open datasets and benchmarks for detecting synthetic images and video [129].

While the goal of fairness for privacy is preventing the imputation of sensitive data, the goal of fairness for censorship is to reveal the unfair prevention of expression. This duality could surface important technical connections between these fields. In terms of social impact, many people around the world outside of the queer community are negatively affected by censorship. Further research in fairness for censorship could have far-reaching benefit across technical fields, social groups and borders.

## 2.3 Language

Language encodes and represents our way of thinking and communicating about the world. There is a long history of oppressive language being weaponised against the queer community [117, 155], highlighting the need for developing fair and inclusive language models.

Inclusive language [163] extends beyond the mere avoidance of derogatory terms, as there are many ways in which harmful stereotypes can surface. For example, the phrase ‘That’s so gay’ [39] equates queerness with badness. Using the term ‘sexual preference’ rather than ‘sexual orientation’ can imply that sexual orientation is a volitional choice, rather than an intrinsic part of one’s identity. Assuming one’s gender identity, without asking, is harmful to the trans community as it risks misgendering people. This can manifest in the careless use of assumed pronouns, without knowledge of an individual’s identification and requested pronouns. Reinforcing binary and traditional gender expression stereotypes, regardless of intent, can have adverse consequences. The use of gender-neutral pronouns has been shown to result in lower bias against women and LGBTQ+ people [151]. To further complicate the matter, words which originated in a derogatory context, such as the label ‘queer’ itself, are often reclaimed by the community in an act of resistance. This historical precedent suggests that AI systems must be able to adapt to the evolution of natural language and avoid censoring language based solely on its adjacency to the queer community.

**2.3.1 Promise.** Natural language processing applications permeate the field of AI. These applications include use cases of general interest like machine translation, speech recognition, sentiment analysis, question answering, chatbots and hate speech detection systems. There is an opportunity to develop language-based AI systems inclusively—to overcome human biases and establish inclusive norms that would facilitate respectful communication with regards to sexual orientation and gender identity [147].

**2.3.2 Risks.** Biases, stereotypes and abusive speech are persistently present in top-performing language models, as a result of their presence in the vast quantities of training data that are needed for model development [44]. Formal frameworks for measuring and ensuring fairness [69, 70, 141] in language are still in nascent stages of development. Thus, for AI systems to avoid reinforcing harmful stereotypes and perpetuating harm to marginalised groups,

research on inclusive language requires more attention. For language systems to be fair, they must be capable of reflecting the contextual nature of human discourse.

## 2.4 Fighting Online Abuse

The ability to safely participate in online platforms is critical for marginalised groups to form a community and find support [104]. However, this is often challenging due to pervasive online abuse [80]. Queer people are frequently targets of internet hate speech, harassment and trolling. This abuse may be directed at the community as a whole or at specific individuals who express their queer identity online. Adolescents are particularly vulnerable to cyberbullying and the associated adverse effects, including depression and suicidal ideation [2]. Automated systems for moderation of online abuse are a possible solution that can protect the psychological safety of the queer community at a global scale.

**2.4.1 Promise.** AI systems could potentially be used to help human moderators flag abusive online content and communication directed at members of marginalised groups, including the queer community [131, 134]. A proof of concept for this application was developed in the Troll Patrol project [6, 50], a collaboration between Amnesty International and Element AI’s former AI for Good team. The Troll Patrol project investigated the application of natural language processing methods for quantifying abuse against women on Twitter. The project revealed concerning patterns of online abuse and highlighted the technological challenges required to develop online abuse detection systems. Recently, similar systems have been applied to tweets directed at the LGBTQ+ community. Machine learning and sentiment analysis were leveraged to predict homophobia in Portuguese tweets, resulting in 89.4% accuracy [125]. Deep learning has also been used to evaluate the level of public support and perception of LGBTQ+ rights following the Supreme Court of India’s verdict regarding the decriminalisation of homosexuality [88].

The ways in which abusive comments are expressed when targeted at the trans community pose some idiosyncratic research challenges. In order to protect the psychological safety of trans people, it is necessary for automated online abuse detection systems to properly recognise acts of misgendering (referring to a trans person with the gender they were assigned at birth) or ‘deadnaming’ (referring to a trans person by the name they had before they transitioned; [146]). These systems have a simultaneous responsibility to ensure that deadnames and other sensitive information are kept private to the user. It is therefore essential for the queer community to play an active role in informing the development of such systems.

**2.4.2 Risks.** Systems developed with the purpose of automatically identifying toxic speech could introduce harms by failing to recognise the context in which speech occurs. Mock impoliteness, for example, helps queer people cope with hostility; the communication style of drag queens in particular is often tailored to be provocative. A recent study [61] demonstrated that an existing toxicity detection system would routinely consider drag queens to be as offensive as white supremacists in their online presence. The system further

specifically associated high levels of toxicity with words like ‘gay’, ‘queer’ and ‘lesbian’.

Another risk in the context of combating online abuse is unintentionally disregarding entire groups through ignorance of intersectional issues. Queer people of colour experience disproportionate exposure to online (and offline) abuse [13], even within the queer community itself. This imbalance stems from intersectional considerations about the ways in which race, class, gender, and other individual characteristics may combine into differential modes of discrimination and privilege (“intersectionality”; [46]). Neglecting intersectionality can lead to disproportionate harms for such sub-communities.

To mitigate these concerns, it is important for the research community to employ an inclusive and participatory approach [107] when compiling training datasets for abusive speech detection. For example, there are homophobic and transphobic slurs with a racialised connotation that should be included in training data for abuse detection systems. Furthermore, methodological improvements may help advance progress. Introducing fairness constraints to model training has demonstrably helped mitigate the bias of cyber-bullying detection systems [60]. Adversarial training can similarly assist by demoting the confounds associated with texts of marginalised groups [165].

## 2.5 Health

The drive towards equitable outcomes in healthcare entails a set of unique challenges for marginalised communities. Queer communities have been disproportionately affected by HIV [143], suffer a higher incidence of sexually-transmitted infections, and are afflicted by elevated rates of substance abuse [160]. Compounding these issues, queer individuals frequently experience difficulties accessing appropriate care [23, 75]. Healthcare professionals often lack appropriate training to best respond to the needs of LGBTQ+ patients [135]. Even in situations where clinicians do have the proper training, patients may be reluctant to reveal their sexual orientation and gender identity, given past experiences with discrimination and stigmatisation.

In recent months, the COVID-19 pandemic has amplified health inequalities [29, 157]. Initial studies during the pandemic have found that LGBTQ+ patients are experiencing poorer self-reported health compared to cisgendered heterosexual peers [120]. The health burden of COVID-19 may be especially severe for queer people of colour, given the substantial barriers they face in accessing healthcare [74].

**2.5.1 Promise.** To this day, the prevalence of HIV among the queer community remains a major challenge. Introducing systems that both reduce the transmission risk and improve care delivery for HIV+ patients will play a critical role in improving health outcomes for queer individuals.

Machine learning presents key opportunities to augment medical treatment decisions [21]. For example, AI may be productively applied to identify the patients most likely to benefit from pre-exposure prophylaxis for HIV. A research team recently developed such a system, which correctly identified 38.6% of future cases of HIV [106]. The researchers noted substantial challenges: model sensitivity on the validation set was 46.4% for men and 0% for

women, highlighting the importance of intersectionality for fair outcomes in healthcare. Machine learning has also been used to predict early virological suppression [22], adherence to anti-retroviral therapy [139], and individual risk of complications such as chronic kidney disease [130] or antiretroviral therapy-induced mitochondrial toxicity [97].

**2.5.2 Risks.** Recent advances in AI in healthcare may lead to widespread increases in welfare. Yet there is a risk that benefits will be unequally distributed—and an additional risk that queer people’s needs will not be properly met by the design of current systems. Information about sexual orientation and gender identity is frequently absent from research datasets. To mitigate the privacy risk for patients and prevent reidentification, HIV status and substance abuse are also routinely omitted from published data. While such practices may be necessary, it is worth recognising the important downstream consequences they have for AI system development in healthcare. It can become impossible to assess fairness and model performance across the omitted dimensions. Moreover, the unobserved data increase the likelihood of reduced predictive performance (since the features are dropped), which itself results in worse health outcomes. The coupled risk of a decrease in performance and an inability to measure it could drastically limit the benefits from AI in healthcare for the queer community, relative to cisgendered heterosexual patients. To prevent the amplification of existing inequities, there is a critical need for targeted fairness research examining the impacts of AI systems in healthcare for queer people.

To help assess the quality of care provided to LGBTQ+ patients, there have been efforts aimed at approximately identifying sexual orientation [24] and gender identity [53] from clinical notes within electronic health record systems. While well intentioned, these machine learning models offer no guarantee that they will only identify patients who have explicitly disclosed their identities to their healthcare providers. These models thus introduce the risk that patients will be outed without their consent. Similar risks arise from models developed to rapidly identify HIV-related social media data [168].

The risk presented by AI healthcare systems could potentially intensify during medical gender affirmation. There are known adverse effects associated with transition treatment [115]. The active involvement of medical professionals with experience in cross-sex hormonal therapy is vital for ensuring the safety of trans people undergoing hormone therapy or surgery. Since cisgendered individuals provide the majority of anonymised patient data used to develop AI systems for personalised healthcare, there will be comparatively fewer cases of trans patients experiencing many medical conditions. This scarcity could have an adverse impact on model performance—there will be an insufficient accounting for the interactions between the hormonal treatment, its adverse effects and potential comorbidities, and other health issues potentially experienced by trans patients.

Framing fairness as a purely technical problem that can be addressed by the mere inclusion of more data or computational adjustments is ethically problematic, especially in high-stakes domains like healthcare [110]. Selection bias and confounding in retrospective data make causal inference particularly hard in this domain.

Counterfactual reasoning may prove key for safely planning interventions aimed at improving health outcomes [127]. It is critical for fairness researchers to engage deeply with both clinicians and patients to ensure that their needs are met and AI systems in health-care are developed and deployed safely and fairly.

## 2.6 Mental Health

Queer people are more susceptible to mental health problems than their heterosexual and cisgender peers, largely as a consequence of the chronically high levels of stress associated with prejudice, stigmatisation and discrimination [108, 113, 114, 152]. As a result, queer communities experience substantial levels of anxiety, depression and suicidal ideation [112]. Compounding these issues, queer people often find it more difficult to ask for help and articulate their distress [111] and face systemic barriers to treatment [128]. A recent LGBTQ+ mental health survey highlighted the shocking extent of issues permeating queer communities [153]: 40% of LGBTQ+ respondents seriously considered attempting suicide in the past twelve months, with more than half of transgender and nonbinary youth having seriously considered suicide; 68% of LGBTQ+ youth reported symptoms of generalised anxiety disorder in the past two weeks, including more than three in four transgender and nonbinary youth; 48% of LGBTQ+ youth reported engaging in self-harm in the past twelve months, including over 60% of transgender and nonbinary youth.

**2.6.1 Promise.** AI systems have the potential to help address the alarming prevalence of suicide in the queer community. Natural language processing could be leveraged to help human operators identify cases of an increased suicide risk more reliably, and respond to them appropriately. These predictions could be based either on traditional data sources such as questionnaires and recorded interactions with mental health support workers, or new data sources including social media and engagement data. The Trevor Project, a prominent American organisation providing crisis intervention and suicide prevention services to LGBTQ+ youth [154], is working on such an initiative. The crisis contact simulator developed by The Trevor Project team has been designed to emulate plausible conversations and interactions between the helpline workers and the callers. The Trevor Project uses this simulator to help train new team members, allowing them to practice their skills. The narrow focus of this application mitigates some of the risks intrinsic to the application of language models, though the effectiveness of the system is still being evaluated. In partnership with Google.org and its research fellows, The Trevor Project also developed an AI system to identify and prioritise community members at high risk while simultaneously increasing outreach to new contacts. The system was designed to relate different types of intake-form responses to downstream diagnosis risk levels. A separate group of researchers developed a language processing system [101] to identify help-seeking conversations on LGBTQ+ support forums, with the aim of helping at-risk individuals manage and overcome their issues.

In other healthcare contexts, reinforcement learning has recently demonstrated potential in steering behavioural interventions [166] and improving health outcomes. Reinforcement learning represents a natural framework for personalised health interventions, since

it can be set up to maximise long-term physical and mental well-being [149]. If equipped with natural language capabilities, such systems might be able to act as personalised mental health assistants empowered to support mental health and escalate situations to human experts in concerning situations.

**2.6.2 Risks.** Substantial risks accompany these applications. Overall, research on any intervention-directed systems should be undertaken in partnership with trained mental health professionals and organisations, given the considerable risks associated with misdiagnosing mental illness (cf. [148]) and exacerbating the vulnerability of those experiencing distress.

The automation of intervention decisions and mental health diagnoses poses a marked risk for the trans community. In most countries, patients must be diagnosed with gender dysphoria—an extensive process with lengthy wait times—before receiving treatments such as hormone therapy or surgery (e.g., [119]). During this process, many transgender individuals experience mistrust and invalidation of their identities from medical professionals who withhold treatment based on rigid or discriminatory view of gender [11]. Automating the diagnosis of gender dysphoria may recapitulate these biases and deprive many transgender patients of access to care.

Mental health information is private and sensitive. While AI systems have the potential to aid mental health workers in identifying at-risk individuals and those who would most likely benefit from intervention, such models may be misused in ways that expose the very people they were designed to support. Such systems could also lead queer communities to be shut out from employment opportunities or to receive higher health insurance premiums. Furthermore, reinforcement learning systems for behavioural interventions will present risks to patients unless many open problems in the field can be resolved, such as safe exploration [66] and reward specification [92]. The development of safe intervention systems that support the mental health of the queer community is likely also contingent on furthering frameworks for sequential fairness [68], to fully account for challenges in measuring and promoting queer ML fairness.

## 2.7 Employment

Queer people often face discrimination both during the hiring process (resulting in reduced job opportunities) and once hired and employed (interfering with engagement, development and well-being; [137]). Non-discrimination laws and practices have had a disparate impact across different communities. Employment nondiscrimination acts in the United States have led to an average increase in the hourly wages of gay men by 2.7% and a decrease in employment of lesbian women by 1.7% [33], suggesting that the impact of AI on employment should be examined through an intersectional lens.

**2.7.1 Promise.** To effectively develop AI systems for hiring, researchers must first attempt to formalise a model of the hiring process. Formalising such models may make it easier to inspect current practices and identify opportunities for removing existing biases (e.g., [100]). Incorporating AI into employment decision processes could potentially prove beneficial if unbiased systems are

developed [73], though this seems difficult at the present moment and carries serious risks.

**2.7.2 Risks.** Machine learning-based decision making systems (e.g., candidate prioritisation systems) developed using historical data could assign lower scores to queer candidates, purely based on historical biases. Prior research has demonstrated that resumes containing items associated with queerness are scored significantly lower by human graders than the same resumes with such items removed [96]. These patterns can be trivially learned and reproduced by resume-parsing machine learning models.

A combination of tools aimed at social media scraping, linguistic analysis, and an analysis of interests and activities could indirectly infringe of candidates' privacy by outing them to their prospective employers without their prior consent. The interest in these tools stems from the community's emphasis on big data approaches, not all of which will have been scientifically verified from the perspective of impact on marginalised groups.

Both hiring and subsequent employment are multi-stage processes of considerable complexity, wherein technical AI tools may be used across multiple stages. Researchers will not design and develop truly fair AI systems by merely focusing on metrics of sub-systems in the process, abstracting away the social context of their application and their interdependence. It is instead necessary to see these as *sociotechnical systems* and evaluate them as such [138].

### 3 SOURCES OF UNOBSERVED CHARACTERISTICS

Most algorithmic fairness studies have made progress because of their focus on *observed* characteristics—commonly, race and legal gender. To be included in training or evaluation data for an algorithm, an attribute must be measured and recorded. Many widely available datasets thus focus on immutable characteristics (such as ethnic group) or characteristics which are recorded and regulated by governments (such as legal gender, monetary income or profession).

In contrast, characteristics like sexual orientation and gender identity are frequently *unobserved* [8, 47, 78]. Multiple factors contribute to this lack of data. In some cases, the plan for data collection fails to incorporate questions on sexual orientation and gender identity—potentially because the data collector did not consider or realise that they are important attributes to record [71]. As a result, researchers may inherit datasets where assessment of sexual orientation and gender identity is *logistically excluded*. In other situations, regardless of the surveyor's intent, the collection of certain personal data may threaten an individual's privacy or their safety. Many countries have legislation that actively discriminates against LGBTQ+ people [76]. Even in nations with hard-won protections for the queer community, cultural bias persists. To shield individuals from this bias and protect their privacy, governments may instate legal protections for sensitive data, including sexual orientation [56]. As a result, such data may be *ethically or legally precluded* for researchers. Finally, as recognised by discursive theories of gender and sexuality, sexual orientation and gender identity are fluid cultural constructs that may change over time and across social contexts [35]. Attempts to categorise, label, and record such

information may be inherently ill-posed [64]. Thus, some characteristics are unobserved because they are *fundamentally unmeasurable*. These inconsistencies in awareness and measurability yield discrepancies and tension in how fairness is applied across different contexts [25].

Studies of race and ethnicity are not immune to these challenges. Race and ethnicity may be subject to legal observability issues in settings where race-based discrimination is a sensitive issue (e.g., hiring). Additionally, the definition of racial and ethnic groups has fluctuated across time and place [65]. This is exemplified by the construction of Hispanic identity in the United States and its inclusion on the National Census, as well as the exclusion of multiracial individuals from many censuses until relatively recently [116]. Though we choose to focus our analysis on queer identity, we note that the observability and measurability of race are also important topics (e.g., [133]).

## 4 AREAS FOR FUTURE RESEARCH

The field of algorithmic fairness in machine learning is rapidly expanding. To date, however, most studies have overlooked the implications of their work for queer people. To include sexual orientation and gender identity in fairness research, it will be necessary to explore new technical approaches and evaluative frameworks. To prevent the risk of AI systems harming the queer community—as well as other marginalised groups whose defining features are similarly unobserved and unmeasurable—fairness research must be expanded.

### 4.1 Expanding Fairness for Queer Identities

Machine learning models cannot be considered fair unless they explicitly factor in and account for fairness towards the LGBTQ+ community. To minimise the risks and harms to queer people worldwide and avoid contributing to ongoing erasures of queer identity, researchers must propose solutions that explicitly account for fairness with respect to the queer community.

The intersectional nature of sexual orientation and gender identity [123] emerges as a recurring theme in our discussions of online abuse, health and employment. These identities cannot be understood without incorporating notions of economic and racial justice. Deployed AI systems may pose divergent risks to different queer subcommunities; AI risks may vary between gay, bisexual, lesbian, transgender and other groups. It is therefore important to apply an appropriate level of granularity to the analysis of fairness for algorithmic issues. Policies can simultaneously improve the position of certain queer groups while adversely affecting others—highlighting the need for an intersectional analysis of queer fairness.

Demographic parity has been the focus of numerous ML fairness studies and seems to closely match people's conceptions of fairness [145]. However, this idea is hard to promote in the context of queer ML fairness. Substantial challenges are posed by the sensitivity of group membership information and its absence from most research datasets, as well as the associated outing risks associated with attempts to automatically derive such information from existing data [24, 53, 168]. Consensually provided self-identification data, if and when available, may only capture a fraction of the community. The resulting biased estimates of queer fairness may



involve high levels of uncertainty [55], though it may be possible to utilise unlabeled data for tightening the bounds [82]. While it is possible to root the analysis in proxy groups [63], there is a risk of incorporating harmful stereotypes in proxy group definitions, potentially resulting in harms of representation [1]. Consequently, most ML fairness solutions developed with a specific notion of demographic parity in mind may be inappropriate for ensuring queer ML fairness.

Individual [51, 84], counterfactual [94], and contrastive [36] fairness present alternative definitions and measurement frameworks that may prove useful for improving ML fairness for queer communities. However, more research is needed to overcome implementation challenges for these frameworks and facilitate their adoption.

A small body of work aims to address fairness for protected groups when the collection of protected attributes is legally precluded (e.g., by privacy and other regulation). Adversarially Reweighted Learning [95] aims to address this issue by relying on measurable covariates of protected characteristics (e.g., zip code as a proxy for race). This approach aims to achieve intersectional fairness by optimising group fairness between all computationally identifiable groups [86, 89]. Distributionally robust optimisation represents an alternative method for preventing disparity amplification, bounding the worst-case risk over groups with unknown group membership by optimising the worst-case risk over an appropriate risk region [67]. These methods have helped establish a link between robustness and fairness, and have drawn attention to the synergistic benefits of considering the relationship between fairness and ML generalisation [45]. Other adversarial approaches have also been proposed for improving counterfactual fairness, and by operating in continuous settings, have been shown to be a better fit for protected characteristics that are hard to enumerate [62].

Fairness mitigation methods have been shown to be vulnerable to membership inference attacks where the information leak increases disproportionately for underprivileged subgroups [37]. This further highlights the tension between privacy and fairness, a common theme when considering the impact of AI systems of queer communities. It is important to recognise the need for fairness solutions to respect and maintain the privacy of queer individuals and to be implemented in a way that minimises the associated re-identifiability risks. Differentially private fair machine learning [79] could potentially provide such guarantees, simultaneously meeting the requirements of fairness, privacy and accuracy.

Putting a greater emphasis on model explainability may prove crucial for ensuring ethical and fair AI applications in cases when fairness metrics are hard or impossible to reliably compute for queer communities. Understanding how AI systems operate may help identify harmful biases that are likely to have adverse downstream consequences, even if these consequences are hard to quantify accurately. Even in cases when queer fairness can be explicitly measured, there is value in identifying which input features contribute the most to unfair model outcomes [17], in order to better inform mitigation strategies.

It is important to acknowledge the unquestionable cisnormativity of sex and gender categories traditionally used in the AI research literature. The assumption of fixed, binary genders fails to include and properly account for non-binary identities and trans people [87].

Incorporating such biases in the early stages of AI system design poses a substantial risk of harm to queer people. Moving forward, more attention should be directed to address this lacuna.

Creating more-equitable AI systems will prove impossible without listening to those who are at greatest risk. Therefore, it is crucial for the AI community to involve more queer voices in the development of AI systems, ML fairness, and ethics research [10, 126]. For example, the inclusion of queer perspectives might have prevented the development of natural language systems that inadvertently censor content which is wrongly flagged as abusive or inappropriate simply due to its adjacency to queer culture, such as in the example of scoring drag queen language as toxic. Researchers should make efforts to provide a safe space for LGBTQ+ individuals to express their opinions and share their experiences. Queer in AI workshops have recently been organised at the Neural Information Processing Systems conference [3, 4] and the International Conference on Machine Learning [83], providing a valuable opportunity for queer AI researchers to network in a safe environment and discuss research at the intersection of AI and queer identity.

## 4.2 Fairness for Other Unobserved Characteristics

The queer community is not the only marginalised group for which group membership may be unobserved [47]. Religion, disability status, and class are additional examples where fairness is often challenged by observability [85, 132]. Critically, they may also benefit from developments or solutions within queer fairness research. For example, in nations where individuals of certain religious groups are persecuted or subjected to surveillance, privacy is an essential prerequisite for safety. Persecution targeting religious communities may also include censorship or manipulation of information [42]. Even in nations where religious freedoms are legally protected, religious minorities may be subjected to online abuse such as hate speech or fear-mongering stereotypes [12].

Although the nature of the discrimination is different, people with disabilities are also a frequent target of derogatory language on the internet, and are more likely to be harassed, stalked or trolled online, often to the detriment of their mental health [142]. Youth with disabilities more frequently suffer from adverse mental health due to bullying, and people of all ages with physical disabilities are at higher risk for depression [90, 156]. Therefore, individuals with disabilities may benefit from insights on the interaction of unobserved characteristics and mental health. Lower-income and lower-class individuals also suffer from worse mental health, particularly in countries with high economic inequality [102]. Fairness for class and socioeconomic status is also an important consideration for employment, where class bias in hiring limits employee diversity and may prevent economic mobility [93].

Any particular dataset or AI application may instantiate observability difficulties with respect to multiple demographics. This may frequently be the case for disability status and class, for example. Individual fairness—a set of approaches based on the notion of treating similar individuals similarly [51, 84]—could potentially promote fairness across multiple demographics. These approaches entail a handful of challenges, however. The unobserved group memberships cannot be incorporated in the similarity measure.



As a result, the similarity measure used for assessing individual fairness must be designed carefully. To optimise fairness across multiple demographics and better capture the similarity between people on a fine-grained level, similarity measures will likely need to incorporate a large number of proxy features. This would be a marked divergence from the proposed measures in most published work. Counterfactual and contrastive fairness metrics come with their own set of practical implementation challenges.

On the other hand, the approaches aimed at providing worst-case fairness guarantees for groups with unknown group membership [67, 95] apply by definition to any marginalised group. They are also specifically tailored to address the situation of unobserved protected characteristics. Therefore, fairness solutions required to address queer ML fairness are likely to be applicable to other groups as well.

Fairness challenges are institutionally and contextually grounded, and it is important to go beyond purely computational approaches to fully assess the sociotechnical aspects of the technology being deployed. The complexity of these issues preclude any single group from tackling them in their entirety, and a resolution would ultimately require an ecosystem involving a multitude of partnering organisations, jointly monitoring, measuring and reporting fairness of such systems [158].

These issues are only a small sample of the common challenges faced by groups with typically unobserved characteristics. We invite future work to explore the impact of AI from the perspective of such groups. It is important to acknowledge that people with different identities have distinct experiences of marginalisation, stigmatisation and discrimination. However, recognising common patterns of injustice will likely enable the development of techniques that can transfer across communities and enhance fairness for multiple groups. In this way, shared ethical and technical design principles for AI fairness will hopefully result in a more equitable future.

## 5 CONCLUSION

The queer community has surmounted numerous historical challenges and continues to resist oppression in physical and digital spaces around the world. Advances in artificial intelligence represent both a potential aid to this resistance and a risk of exacerbating existing inequalities. This risk should motivate researchers to design and develop AI systems with fairness for queer identities in mind. Systems that attempt to label sexual orientation and gender identity, even for the purpose of fairness, raise technical and ethical challenges regarding observability and measurability.

A new discourse on queer fairness has the potential to identify moral and practical considerations shared across queer communities, as well as concerns specific to particular subpopulations in particular places. By further developing techniques supporting fairness for unobserved characteristics, the machine learning community can support queer communities and other marginalised groups. Broadly, the present work—surveying the ways in which AI may ameliorate or exacerbate issues faced by queer communities—emphasises the need for machine learning practitioners to design systems with fairness and dignity in mind.

## ACKNOWLEDGMENTS

We would like to thank Aliya Ahmed, Dorothy Chou, Ben Coppin, Michelle Dunlop, Thore Graepel, William Isaac, Koray Kavukcuoglu, Guy Scully, and Laura Weidinger for the support and insightful feedback that they provided for this paper.

Any opinions presented in this paper represent the personal views of the authors and do not necessarily reflect the official policies or positions of their organisations.

## REFERENCES

- [1] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: Quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 801–809.
- [2] Roberto L Abreu and Maureen C Kenny. 2018. Cyberbullying and LGBTQ youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma* 11, 1 (2018), 81–97.
- [3] William Agnew, Samy Bengio, Os Keyes, Margaret Mitchell, Shira Mitchell, Vinodkumar Prabhakaran, David Vazquez, and Raphael Gontijo Lopes. 2018. Queer in AI 2018 Workshop. <https://queerai.github.io/QueerInAI/QinAIatNeurIPS.html>. Accessed: 2020-09-10.
- [4] William Agnew, Natalia Bilenko, and Raphael Gontijo Lopes. 2019. Queer in AI 2019 Workshop. <https://sites.google.com/corp/view/queer-in-ai/neurips-2019/>. Accessed: 2020-09-10.
- [5] Mohammad Al-Rubaie and J Morris Chang. 2019. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy* 17, 2 (2019), 49–58.
- [6] Amnesty International. n.d. Troll Patrol. <https://decoders.amnesty.org/projects/troll-patrol>. Accessed: 2020-09-10.
- [7] Amnesty International Canada. 2015. LGBTI Rights. <https://www.amnesty.ca/our-work/issues/lgbti-rights>. Accessed: 2020-09-10.
- [8] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2020. ‘What we can’t measure, we can’t understand’: Challenges to demographic data procurement in the pursuit of fairness. *arXiv preprint arXiv:2011.02282* (2020), 1–12.
- [9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [10] Sherry R Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of Planners* 35, 4 (1969), 216–224.
- [11] Florence Ashley. 2019. Gatekeeping hormone replacement therapy for transgender patients is dehumanising. *Journal of Medical Ethics* 45, 7 (2019), 480–482.
- [12] Imran Awan. 2014. Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet* 6, 2 (2014), 133–150.
- [13] Kimberly F Balsam, Yamile Molina, Blair Beadnell, Jane Simoni, and Karina Walters. 2011. Measuring multiple minority stress: the LGBT People of Color Microaggressions Scale. *Cultural Diversity and Ethnic Minority Psychology* 17, 2 (2011), 163.
- [14] David Bamman, Brendan O’Connor, and Noah Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday* (2012).
- [15] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. <http://www.fairmlbook.org>.
- [16] Tom L Beauchamp and James F Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press, USA.
- [17] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020), 1–15.
- [18] Joel Bellenson. 2019. 122 Shades of Gray. <https://www.geneplaza.com/app-store/72/preview>. Accessed: 2020-09-10.
- [19] Nikhil Bhattasali and Esha Maiti. 2015. Machine ‘gaydar’: Using Facebook profiles to predict sexual orientation.
- [20] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.
- [21] Kuteesa R Bisaso, Godwin T Anguzu, Susan A Karungi, Agnes Kiragga, and Barbara Castelnuovo. 2017. A survey of machine learning applications in HIV clinical research and care. *Computers in Biology and Medicine* 91 (2017), 366–371.
- [22] Kuteesa R Bisaso, Susan A Karungi, Agnes Kiragga, Jackson K Mukonzo, and Barbara Castelnuovo. 2018. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Medical Informatics and Decision Making* 18, 1 (2018), 1–10.
- [23] Raphaël Bize, Erika Volkmar, Sylvie Berrut, Denise Medico, Hugues Balthasar, Patrick Bodenmann, and Harvey J Makadon. 2011. Access to quality primary care for LGBT people. *Revue Medicale Suisse* 7, 307 (2011), 1712–1717.

- [24] Ragnhildur I Bjarnadóttir, Walter Bockting, Sunmoo Yoon, and Dawn W Dowling. 2019. Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study. *CIN: Computers, Informatics, Nursing* 37, 4 (2019), 213–221.
- [25] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 492–500.
- [26] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.
- [27] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.
- [28] Michael J Bosia, Sandra M McEvoy, and Momin Rahman. 2020. *The Oxford Handbook of Global LGBT and Sexual Diversity Politics*. Oxford University Press.
- [29] Lisa Bowleg. 2020. We're not all in this together: On COVID-19, intersectionality, and structural inequality. *American Journal of Public Health* 110, 7 (2020), 917.
- [30] Brook. n.d.. Gender: A few definitions. <https://www.brook.org.uk/your-life/gender-a-few-definitions/>. Accessed: 2020-10-01.
- [31] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020), 1–75.
- [32] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
- [33] Ian Burn. 2018. Not all laws are created equal: Legal differences in state non-discrimination laws and the impact of LGBT employment protections. *Journal of Labor Research* 39, 4 (2018), 462–497.
- [34] Joseph Burridge. 2004. 'I am not homophobic but...': Disclaiming in discourse resisting repeal of Section 28. *Sexualities* 7, 3 (2004), 327–344.
- [35] Judith Butler. 2011. *Bodies That Matter: On the Discursive Limits of Sex*. Taylor & Francis.
- [36] Tapabrata Chakraborti, Arijit Patra, and J Alison Noble. 2020. Contrastive fairness in machine learning. *IEEE Letters of the Computer Society* 3, 2 (2020), 38–41.
- [37] Hongyan Chang and Reza Shokri. 2020. On the privacy risks of algorithmic fairness. *arXiv preprint arXiv:2011.03731* (2020), 1–14.
- [38] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.
- [39] Jill M Chonody, Scott Edward Rutledge, and Scott Smith. 2012. 'That's so gay': Language use and antigay bias among heterosexual college students. *Journal of Gay & Lesbian Social Services* 24, 3 (2012), 241–259.
- [40] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.
- [41] Jennifer Cobbe. 2019. Algorithmic censorship on social platforms: Power, legitimacy, and resistance. *Legitimacy, and Resistance* (2019).
- [42] Sarah Cook. 2017. *The Battle for China's Spirit: Religious Revival, Repression, and Resistance under Xi Jinping*. Rowman & Littlefield.
- [43] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018), 1–25.
- [44] Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence* 1, 11 (2019), 495–496.
- [45] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2020. Exchanging lessons between algorithmic fairness and domain generalization. *arXiv preprint arXiv:2010.07249* (2020).
- [46] Kimberlé Crenshaw. 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.
- [47] J Crocker, B Major, and C Steele. 1998. Social stigma. In *The Handbook of Social Psychology*, Daniel Todd Gilbert, Susan T Fiske, and Gardner Lindzey (Eds.), Vol. 1. Oxford University Press.
- [48] Natasha Culzac. 2014. Egypt's police 'using social media and apps like Grindr to trap gay people'. <https://www.independent.co.uk/news/world/africa/egypt-s-police-using-social-media-and-apps-grindr-trap-gay-people-9738515.html>.
- [49] Christina DeJong and Eric Long. 2014. The death penalty as genocide: The persecution of 'homosexuals' in Uganda. In *Handbook of LGBT Communities, Crime, and Justice*. Springer, 339–362.
- [50] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. 2019. A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. *arXiv preprint arXiv:1902.03093* (2019), 1–13.
- [51] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.
- [52] Wayne R Dynes. 2014. *The Homophobic Mind*. Lulu.com.
- [53] Jesse M Ehrenfeld, Keanan Gabriel Gottlieb, Lauren Brittany Beach, Shelby E Monahan, and Daniel Fabbri. 2019. Development of a natural language processing algorithm to identify and evaluate transgender patients in electronic health record systems. *Ethnicity & Disease* 29, Suppl 2 (2019), 441.
- [54] Tom Embury-Dennis. 2020. Bullied and blackmailed: Gay men in Morocco falling victims to outing campaign sparked by Instagram model. <https://www.independent.co.uk/news/world/africa/gay-men-morocco-dating-apps-grindr-instagram-sofia-taloni-a9486386.html>. Accessed: 2020-09-10.
- [55] Kavin Ethayarajh. 2020. Is your classifier actually biased? Measuring fairness under uncertainty with Bernstein bounds. *arXiv preprint arXiv:2004.12332* (2020), 1–6.
- [56] European Commission. n.d.. What personal data is considered sensitive? [https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en). Accessed: 2020-10-07.
- [57] Drew Fudenberg and David K Levine. 2012. Fairness, risk preferences and independence: Impossibility theorems. *Journal of Economic Behavior & Organization* 81, 2 (2012), 606–612.
- [58] Andrea Ganna, Karin JH Verweij, Michel G Nivard, Robert Maier, Robbee Wedow, Alexander S Busch, Abdel Abdellaoui, Shengru Guo, J Fah Sathirapongsasuti, Paul Lichtenstein, et al. 2019. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* 365, 6456 (2019), eaat7693.
- [59] Andrew Gelman, G Marrson, and Daniel Simpson. 2018. Gaydar and the fallacy of objective measurement. Unpublished manuscript. Retrieved from <http://www.stat.columbia.edu/gelman/research/unpublished/gaydar2.pdf>.
- [60] Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *arXiv preprint arXiv:2005.06625* (2020), 1–11.
- [61] Alessandra Gomes, Dennys Antonialli, and Thiago Dias Oliva. 2019. Drag queens and artificial intelligence: Should computers decide what is 'toxic' on the internet? <https://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/>. Accessed: 2020-09-10.
- [62] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2020. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122* (2020), 1–11.
- [63] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018), 1–12.
- [64] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [65] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.
- [66] Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft. 2008. Safe exploration for reinforcement learning. In *ESANN*. 143–148.
- [67] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
- [68] Hoda Heidari and Andreas Krause. 2018. Preventing disparate treatment in sequential decision making. In *International Joint Conference on Artificial Intelligence*. 2248–2254.
- [69] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275* (2020), 1–29.
- [70] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020), 1–27.
- [71] Gregory M Herek, Douglas C Kimmel, Hortensia Amaro, and Gary B Melton. 1991. Avoiding heterosexual bias in psychological research. *American Psychologist* 46, 9 (1991), 957.
- [72] Leora Hoshall. 2012. Afraid of who you are: No promo homo laws in public school sex education. *Tex. J. Women & L.* 22 (2012), 219.
- [73] Kimberly A Houser. 2019. Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.* 22 (2019), 290.
- [74] Ning Hsieh and Matt Ruther. 2016. Sexual minority health and health risk factors: Intersection effects of gender, race, and sexual identity. *American Journal of Preventive Medicine* 50, 6 (2016), 746–755.
- [75] Human Rights Watch. 2018. US: LGBT people face healthcare barriers. <https://www.hrw.org/news/2018/07/23/us-lgbt-people-face-healthcare-barriers>. Accessed: 2020-09-10.

- [76] Human Rights Watch. n.d.. Maps of anti-LGBT laws country by country. [http://internap.hrw.org/features/features/lgbt\\_laws/](http://internap.hrw.org/features/features/lgbt_laws/). Accessed: 2020-10-06.
- [77] Intertech LGBT+ Diversity Forum. n.d.. Intertech LGBT+ Diversity Forum. <https://intertechlgbtinterests.me/>. Accessed: 2020-10-07.
- [78] Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness. *arXiv preprint arXiv:1912.05511* (2019), 1–11.
- [79] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. 2019. Differentially private fair learning. In *International Conference on Machine Learning*. PMLR, 3000–3008.
- [80] Emma A Jane. 2020. Online abuse and harassment. *The International Encyclopedia of Gender, Media, and Communication* (2020), 1–16.
- [81] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*. 1895–1912.
- [82] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference. *Advances in Neural Information Processing Systems* 33 (2020).
- [83] Ti John, William Agnew, Alex Markham, Anja Meunier, Manu Saraswat, Andrew McNamara, and Raphael Gontijo Lopes. 2020. Queer in AI 2020 Workshop. <https://sites.google.com/corp/view/queer-in-ai/icml-2020>. Accessed: 2020-09-10.
- [84] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660* (2019), 1–33.
- [85] Shanna K Kattari, Miranda Olzman, and Michele D Hanna. 2018. 'You look fine!' Ableist experiences by people with invisible disabilities. *Affilia* 33, 4 (2018), 477–492.
- [86] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.
- [87] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.
- [88] Aparup Khatua, Erik Cambria, Kuntal Ghosh, Nabendu Chaki, and Apalak Khatua. 2019. Tweeting in support of LGBT? A deep learning approach. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 342–345.
- [89] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems* 31 (2018), 4842–4852.
- [90] Tania King, Zoe Aitken, Allison Milner, Eric Emerson, Naomi Priest, Amalia Karahalios, Anne Kavanagh, and Tony Blakely. 2018. To what extent is the association between disability and mental health in adolescents mediated by bullying? A causal mediation analysis. *International Journal of Epidemiology* 47, 5 (2018), 1402–1413.
- [91] Alexander Kondakov. 2019. The censorship 'propaganda' legislation in Russia. *State-Sponsored Homophobia* (2019).
- [92] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification gaming: The flip side of AI ingenuity. <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity>.
- [93] Michael W Kraus, Brittany Torrez, Jun Won Park, and Fariba Ghayebi. 2019. Evidence for the reproduction of social class in brief speech. *Proceedings of the National Academy of Sciences* 116, 46 (2019), 22998–23003.
- [94] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
- [95] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. 2020. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114* (2020), 1–15.
- [96] Victoria R LeCroy and Joshua S Rodefer. 2019. The influence of job candidate LGBT association on hiring decisions. *North American Journal of Psychology* 21, 2 (2019).
- [97] Jong Soo Lee, Elijah Paintsil, Vivek Gopalakrishnan, and Musie Ghebremichael. 2019. A comparison of machine learning techniques for classification of HIV patients with antiretroviral therapy-induced mitochondrial toxicity from those without mitochondrial toxicity. *BMC Medical Research Methodology* 19, 1 (2019), 1–10.
- [98] Lesbians Who Tech. n.d.. Lesbians Who Tech. <https://lesbianswhotech.org/>. Accessed: 2020-10-07.
- [99] LGBT Technology Institute. n.d.. LGBT Technology Institute. <https://www.lgbttech.org/>. Accessed: 2020-10-07.
- [100] Danielle Li, Lindsey R Raymond, and Peter Bergman. 2020. *Hiring as exploration*. Technical Report. National Bureau of Economic Research.
- [101] Chen Liang, Dena Abbott, Y Alicia Hong, Mahboubeh Madadi, and Amelia White. 2019. Clustering help-seeking behaviors in LGBT online communities: A prospective trial. In *International Conference on Human-Computer Interaction*. Springer, 345–355.
- [102] William M Liu and Saba R Ali. 2008. Social class and classism: Understanding the psychological impact of poverty and inequality. *Handbook of Counseling Psychology* 4 (2008), 159–175.
- [103] Yujia Liu, Weiming Zhang, and Nenghai Yu. 2017. Protecting privacy in shared photos via adversarial examples based stealth. *Security and Communication Networks* 2017 (2017).
- [104] Yi-Ling Liu. 2020. How a dating app helped a generation of Chinese come out of the closet. <https://www.nytimes.com/2020/03/05/magazine/blued-china-gay-dating-app.html>.
- [105] Shen Lu and Katie Hunt. 2016. China bans same-sex romance from TV screens. <http://www.https://edition.cnn.com/2016/03/03/asia/china-bans-same-sex-dramas/index.html>. Accessed: 2020-09-10.
- [106] Julia L Marcus, Leo B Hurley, Douglas S Krakower, Stacey Alexeeff, Michael J Silverberg, and Jonathan E Volk. 2019. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: A modelling study. *The Lancet HIV* 6, 10 (2019), e688–e695.
- [107] Donald Martin Jr., Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020), 1–6.
- [108] Vickie M Mays and Susan D Cochran. 2001. Mental health correlates of perceived discrimination among lesbian, gay, and bisexual adults in the United States. *American Journal of Public Health* 91, 11 (2001), 1869–1876.
- [109] Joseph McCormick. 2015. WARNING: These Grindr profiles are actually robots trying to steal your info. <https://www.pinknews.co.uk/2015/08/11/warning-these-grindr-profiles-are-actually-robots-trying-to-steal-your-info/>. Accessed: 2020-09-10.
- [110] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. 2020. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* 2, 5 (2020), e221–e223.
- [111] Elizabeth McDermott. 2015. Asking for help online: Lesbian, gay, bisexual and trans youth, self-harm and articulating the 'failed' self. *Health* 19, 6 (2015), 561–577.
- [112] Mental Health Foundation. 2020. Mental health statistics: LGBT people. <https://www.mentalhealth.org.uk/statistics/mental-health-statistics-lgbt-people>. Accessed: 2020-09-10.
- [113] Ilan H Meyer. 1995. Minority stress and mental health in gay men. *Journal of Health and Social Behavior* (1995), 38–56.
- [114] Ilan H Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin* 129, 5 (2003), 674.
- [115] Eva Moore, Amy Wisniewski, and Adrian Dobs. 2003. Endocrine treatment of transsexual people: A review of treatment regimens, outcomes, and adverse effects. *The Journal of Clinical Endocrinology & Metabolism* 88, 8 (2003), 3467–3473.
- [116] G Cristina Mora. 2014. *Making Hispanics: How activists, bureaucrats, and media constructed a new American*. University of Chicago Press.
- [117] Kevin L Nadal, Marie-Anne Issa, Jayleen Leon, Vanessa Meterko, Michelle Wide-man, and Yinglee Wong. 2011. Sexual orientation microaggressions: 'Death by a thousand cuts' for lesbian, gay, and bisexual youth. *Journal of LGBT Youth* 8, 3 (2011), 234–259.
- [118] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 634–646.
- [119] National Health Service. n.d.. Gender dysphoria. <https://www.nhs.uk/conditions/gender-dysphoria/>. Accessed: 2020-09-10.
- [120] Kathryn O'Neill. 2020. Health vulnerabilities to COVID-19 among LGBT adults in California. <https://escholarship.org/uc/item/8hc4z2gb>. Accessed: 2020-09-10.
- [121] Nancy Ordover. 2003. *American Eugenics: Race, Queer Anatomy, and the Science of Nationalism*. U of Minnesota Press.
- [122] Out in Tech. n.d.. Out in Tech. <https://outintech.com/>. Accessed: 2020-10-07.
- [123] Mike C Parent, Cirleen DeBlanc, and Bonnie Moradi. 2013. Approaches to research on intersectionality: Perspectives on gender, LGBT, and racial/ethnic identities. *Sex Roles* 68, 11-12 (2013), 639–645.
- [124] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018), 1–6.
- [125] Vinicius Gomes Pereira. 2018. *Using supervised machine learning and sentiment analysis techniques to predict homophobia in Portuguese tweets*. Ph.D. Dissertation. Fundação Getúlio Vargas.
- [126] Adam Poulsen, Eduard Fosch-Villaronga, and Roger Andre Søraa. 2020. Queering machines. *Nature Machine Intelligence* 2, 3 (2020), 152–152.
- [127] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.
- [128] Meghan Romanelli and Kimberly D Hudson. 2017. Individual and systemic barriers to health care: Perspectives of lesbian, gay, bisexual, and transgender adults. *American Journal of Orthopsychiatry* 87, 6 (2017), 714.

- [129] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1–11.
- [130] Jan A Roth, Gorjan Radevski, Catia Marzolini, Andri Rauch, Huldrych F Günthard, Roger D Kouyos, Christoph A Fux, Alexandra U Scherrer, Alexandra Calmy, Matthias Cavassini, et al. 2020. Cohort-derived machine learning models for individual prediction of chronic kidney disease in people living with Human Immunodeficiency Virus: A prospective multicenter cohort study. *The Journal of Infectious Diseases* (2020).
- [131] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2019. HateMonitors: Language agnostic abuse detection in social media. *arXiv preprint arXiv:1909.12642* (2019), 1–8.
- [132] Maria C Sanchez and Linda Schlossberg. 2001. *Passing: Identity and Interpretation in Sexuality, Race, and Religion*. Vol. 29. NYU Press.
- [133] Morgan Klaus Scheuerman, Kandra Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–35.
- [134] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 1–10.
- [135] Jason S Schneider, Vincent MB Silenzio, and Laura Erickson-Schroth. 2019. *The GLMA Handbook on LGBT Health*. ABC-CLIO.
- [136] Dominic Scicchitano. 2019. The 'real' Chechen man: Conceptions of religion, nature, and gender and the persecution of sexual minorities in postwar Chechnya. *Journal of Homosexuality* (2019), 1–18.
- [137] Brad Sears and Christy Mallory. 2011. Documented evidence of employment discrimination & its effects on LGBT people.
- [138] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.
- [139] John Semerdjian, Konstantinos Lykopoulos, Andrew Maas, Morgan Harrell, Julie Priest, Pedro Eitz-Ferrer, Connor Wyand, and Andrew Zolopa. 2018. Supervised machine learning to predict HIV outcomes using electronic health record and insurance claims data. In *AIDS 2018 Conference*.
- [140] Katherine Sender. 2018. The gay market is dead, long live the gay market: From identity to algorithm in predicting consumer behavior. *Advertising & Society Quarterly* 18, 4 (2018).
- [141] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268* (2020), 1–16.
- [142] Mark Sherry. 2019. Disablist hate speech online. *Disability Hate Speech: Social, Cultural and Political Contexts* (2019).
- [143] Sonia Singh, Ruiguang Song, Anna Satcher Johnson, Eugene McCray, and H Irene Hall. 2018. HIV incidence, prevalence, and undiagnosed infections in US men who have sex with men. *Annals of Internal Medicine* 168, 10 (2018), 685–694.
- [144] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. 2019. Privacy-preserving adversarial representation learning in ASR: Reality or illusion?. In *20th Annual Conference of the International Speech Communication Association*.
- [145] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.
- [146] Stonewall. n.d.. The truth about trans. <https://www.stonewall.org.uk/truth-about-trans>. Accessed: 2021-04-19.
- [147] Yolande Strengers, Lizhen Qu, Qionghai Xu, and Jarrod Knibbe. 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*. 1–14.
- [148] Derek H Suite, Robert La Bril, Annelie Primm, and Phyllis Harrison-Ross. 2007. Beyond misdiagnosis, misunderstanding and mistrust: Relevance of the historical perspective in the medical and mental health treatment of people of color. *Journal of the National Medical Association* 99, 8 (2007), 879.
- [149] Seyed Amin Tabatabaei, Mark Hoogendoorn, and Aart van Halteren. 2018. Narrowing reinforcement learning: Overcoming the cold start problem for personalized health interventions. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 312–327.
- [150] Rima S Tanash, Zhouhan Chen, Tanmay Thakur, Dan S Wallach, and Devika Subramanian. 2015. Known unknowns: An analysis of Twitter censorship in Turkey. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*. 11–20.
- [151] Margit Tavits and Efrén O Pérez. 2019. Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences* 116, 34 (2019), 16781–16786.
- [152] Elliot A Tebbe and Bonnie Moradi. 2016. Suicide risk in trans populations: An application of minority stress theory. *Journal of Counseling Psychology* 63, 5 (2016), 520.
- [153] The Trevor Project. 2020. National survey on LGBTQ youth mental health 2020. <https://www.thetrevorproject.org/survey-2020/>. Accessed: 2020-09-10.
- [154] The Trevor Project. n.d.. The Trevor Project. <https://www.thetrevorproject.org/>. Accessed: 2020-09-10.
- [155] Crispin Thurlow. 2001. Naming the "outsider within": Homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *Journal of Adolescence* 24, 1 (2001), 25–38.
- [156] R Jay Turner and Samuel Noh. 1988. Physical disability and depression: A longitudinal analysis. *Journal of Health and Social Behavior* (1988), 23–37.
- [157] Aaron van Dorn, Rebecca E Cooney, and Miriam I Sabin. 2020. COVID-19 exacerbating inequalities in the US. *Lancet* 395, 10232 (2020), 1243.
- [158] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 1–17.
- [159] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience* 2018 (2018).
- [160] Barbara C Wallace and Erik Santacruz. 2017. Addictions and substance abuse in the LGBT community: New approaches. *LGBT Psychology and Mental Health: Emerging Research and Advances* (2017), 153–175.
- [161] Yuanyuan Wang, Zhishan Hu, Ke Peng, Ying Xin, Yuan Yang, Jack Drescher, and Runsen Chen. 2019. Discrimination against LGBT populations in China. *The Lancet Public Health* 4, 9 (2019), e440–e441.
- [162] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114, 2 (2018), 246.
- [163] Michael Weinberg. 2009. LGBT-inclusive language. *English Journal* 98, 4 (2009), 50.
- [164] Gregor Wolbring. 2001. Where do we draw the line? Surviving eugenics in a technological world. *Disability and the Life Course: Global Perspectives* (2001), 38–49.
- [165] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246* (2020), 1–8.
- [166] Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. 2017. Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *Journal of Medical Internet Research* 19, 10 (2017), e338.
- [167] Jillian York. 2015. Privatising censorship online. *Global Information Society Watch 2015: Sexual rights and the internet* (2015), 26–29.
- [168] Sean D Young, Wenchao Yu, and Wei Wang. 2017. Toward automating HIV identification: Machine learning for rapid identification of HIV-related social media data. *Journal of Acquired Immune Deficiency Syndromes* 74, Suppl 2 (2017), S128.
- [169] Jiaming Zhang, Jitao Sang, Xian Zhao, Xiaowen Huang, Yanfeng Sun, and Yongli Hu. 2020. Adversarial privacy-preserving filter. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1423–1431.