Gendering algorithms in social media

Eduard Fosch-Villaronga
eLaw Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands
e.fosch.villaronga@law.leidenuniv.nl

Adam Poulsen
School of Computing and Mathematics, Charles Sturt University, Australia
apoulsen@csu.edu.au

Roger A. Søraa
Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway
roger.soraa@ntnu.no

Bart Custers
eLaw Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands
b.h.m.custers@law.leidenuniv.nl

ABSTRACT

Social media platforms employ inferential analytics methods to guess user preferences and may include sensitive attributes such as race, gender, sexual orientation, and political opinions. These methods are often opaque, but they can have significant effects such as predicting behaviors for marketing purposes, influencing behavior for profit, serving attention economics, and reinforcing existing biases such as gender stereotyping. Although two international human rights treaties include express obligations relating to harmful and wrongful stereotyping, these stereotypes persist both online and offline, and platforms often appear to fail to understand that gender is not merely a binary of being a 'man' or a 'woman,' but is socially constructed. Our study investigates the impact of algorithmic bias on inadvertent privacy violations and the reinforcement of social prejudices of gender and sexuality through a multidisciplinary perspective including legal, computer science, and queer media viewpoints. We conducted an online survey to understand whether and how Twitter inferred the gender of users. Beyond Twitter's binary understanding of gender and the inevitability of the gender inference as part of Twitter's personalization trade-off, the results show that Twitter misgendered users in nearly 20% of the cases (N=109). Although not apparently correlated, only 8% of the straight male respondents were misgendered, compared to 25% of gay men and 16% of straight women. Our contribution shows how the lack of attention to gender in gender classifiers exacerbates existing biases and affects marginalized communities. With our paper, we hope to promote the online account for privacy, diversity, and inclusion and advocate for the freedom of identity that everyone should have online and offline.

Keywords

Gender; Twitter; Inference; Gender Classifier; Privacy; Algorithmic Bias; Discrimination; LGBTQAI+; Gender Stereotyping; Social Media

1. INTRODUCTION

Online and social media platform providers use users' traits, including name, age, and gender, to improve user experience and personalize online behavioral advertising. By knowing users' characteristics, corporations can target or exclude certain groups more efficiently, tailor their services to users, and increase the time they spend on the platform [61]. In such a way, profiling makes marketing more precise and effective. However, a growing concern is the increasing use of opaque inferential analytics that reveal sensitive user attributes that serve attention economics [15] and that may reinforce existing biases which, although not explicit, can be very influential [10, 14].

A recurrent bias is gender stereotyping. Gender stereotyping "refers to the practice of ascribing to an individual 'woman' or 'man' specific attributes, characteristics, or roles by reason only of their membership in the social group of 'women or men'" [59]. However, gender stereotyping is a complex process that, although grounded in strong beliefs of what a gender is and should be, is both used and understood in a too simplistic manner. For instance, gay men are hyper-sexualized in e.g., the masculine promiscuity stereotype, or feminized, e.g., gay men who are perceived to be feminine fall into traditional female stereotypes. Two international human rights treaties include express obligations relating to harmful and wrongful stereotyping. Art. 5 of the Convention on the Elimination of All Forms of Discrimination against Women mandates States Parties to "take all appropriate measures to modify the social and cultural patterns of conduct of men and women, to achieve the elimination of prejudices and customary and all other practices which are based on the idea of the inferiority or the superiority of either of the sexes or stereotyped roles for men and women" [1]. Art. 8(1)(b) of the Convention on the Rights of Persons with Disabilities stresses that "States Parties undertake to adopt immediate, effective and appropriate measures to combat stereotypes, prejudices and harmful practices relating to persons with disabilities, including those based on sex and age, in all areas of life" [2]. However, these stereotypes are preserved both online and offline [22, 26]; platforms often appear to fail to grasp that gender is not limited to the simple binary of being solely a "man" or a "woman," but socially constructed [8].

Given that gender stereotypes persist online, and that the social media platform Twitter infers gender from a wide variety of sources,[1] we address the research question (RQ): *How accurate are Twitter's inferences of its users' gender identities?* Addressing this RQ brings into view concerns of discrimination, misgendering, and exacerbation of existing biases that online platforms persist in replicating that has already been highlighted by existing literature [23, 33]. Our goal is to investigate misgendering on Twitter and

[1] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases

illustrate the impact of algorithmic bias on inadvertent privacy violations and how such biases reinforce social prejudices of gender and sexuality through a multidisciplinary perspective including legal, computer science, and queer media-studies viewpoints.

The reason behind our contribution lies in the idea that gender is a co-shaped, changing part of human identity tied into the socio-materiality of gendered relations often treated as a binary dichotomy. For instance, trans and non-binary users have recently claimed that they are being misgendered on Twitter because the categories "female" and "male" do not match who they are [16]. Second, platform providers no longer have to learn sensitive details about a particular user or correctly group users into categories for advertising to be effective, as advertising has a high tolerance for classification errors [62]. Nonetheless, not considering a broader understanding of gender in platforms can be socially harmful and costly, as technology usage and implementation may lead to further exacerbation of existing biases, including those relating to gender, race, and minorities [5, 24, 54].

In Section 2 of this article, we provide background information on inferential analytics to elucidate how companies infer specific user attributes, including gender, and how these techniques may harm users' rights. In Section 3, we explain the methods for this study, and in Section 4 we introduce the results. Our findings suggest that Twitter's binary understanding of gender excludes those not fitting the category "male" and "female." The results also show that inferring gender is part of Twitter's personalization trade-off and misgenders users in nearly 20% of the cases. Out of these cases, LGBTQIA+ individuals and straight women were misgendered more frequently than straight men. In Section 5 we discuss the lack of diversity in social media platforms and the role designers play in accounting for inclusivity and diversity. We conclude by presenting our future work, which includes a more extensive and refined survey to investigate this issue and the user's impressions further.

2. GENDERING ALGORITHMS

2.1 Profiling, inference analytics, and discrimination

Profiling techniques like regression, classification, or clustering mainly ascribe properties to people [9]. These methods infer distinct people's traits from different inputs of data, originating either from the person themself (i.e., predicting recidivism based on someone's criminal record) or others (i.e., others who ordered these shoes also like these shoes). Organizations use inferential analytics to induce user preferences using sensitive attributes such as race, gender, sexual orientation, political interests, and opinions [30, 56, 63]. These techniques can predict behaviors for marketing purposes and influence behavior for profit [68]. A critical feature of inferential analytics is that companies infer information from data not directly or indirectly provided by data subjects [14]. These inferences may be precise (like inferring age from the date of birth) or estimates (like inferring emotional states, e.g. happiness, or even intelligence from Facebook likes) [35]. In this way, data analytics can predict qualities that a data subject may not want to disclose and attributes that a data subject does not even know about themselves and ascribe them to an individual person.

One of the parameters used to infer attributes from people is the "like" button on many social media platforms [53]. In other words, what users like online tells something about who they are, such as their income [42] with a high degree of accuracy. Gender can also be inferred from Facebook likes with very high accuracy [35]. With approximately 250 Facebook likes, gender could be predicted with accuracy rates of 93%. Although this may seem like a high number, gender could be predicted with accuracy rates of about 70% when using only five Facebook likes. Moreover, when using only one single Facebook like, the accuracy rates were approximately 60% for gender predictions. According to Kosinski *et al.*, predictions for homosexuality were about 88% accurate for gays and 75% for lesbians, and predictions on being single versus in a relationship were about 67% accurate [35]– showing the complexity of inferring gender and sexuality through likes alone.

Inferential analytics may have some benefits. For instance, it can be a tool to fill gaps in fragmentary datasets or check the accuracy of available data by matching inferred data with the contested data. In this way, datasets enriched with many inferred attributes are likely to have higher levels of completeness and precision. In big data analytics, completeness and correctness of data is not a strict condition but can contribute to getting more well-defined and reliable results. Companies can identify that a particular customer prefers to consume video instead of text content, or is interested in learning about particular topics, like travel, fashion, or food. Companies use this information to personalize the user experience to fit the preferences of that particular individual.

However, inferential analytics has some drawbacks. When people's attributes are predicted, privacy is at stake, especially if people did not want to disclose specific personal information. Furthermore, these inferences may contain errors, leading to biased and unfair decisions and may lead to self-fulfilling prophecies [13]. These effects may amplify inequality, undermine democracy, lead to opinion echo chambers, and further push people into categories that are hard to break out of [49].

Machine learning and data mining tools can be developed so that they do not grant discriminating patterns such as gender stereotypes or profiles, a practice called discrimination-aware data mining [31]. The underlying idea is not to limit the data input (such as gender data), but to prevent the algorithms from yielding gender-based patterns, since not using gender data may still allow for predicting gender and thus result in indirect discrimination (discrimination by proxy). Focusing on the algorithms' design can prevent this when using gender in the development of data-driven decision models [67].

2.2 Gender inferences

Gender classification systems (GCS) are trained using a training dataset (or corpus) of structured and labeled data. These labels categorize data, and the features within, as either masculine or feminine [51]. Training a GCS builds a classification algorithm (or classifier) that categorizes features—such as body movements, physiological and behavioral characteristics, and facial features [51]—found in new data by comparing it to labeled features in the dataset. A GCS uses a feature extraction algorithm, classifier, and a dataset to make an inference [39].

Classifiers are trained in machine learning models. Exemplary models include neural networks [51], K-nearest neighbor [34], support vector machine [38], and Adaboost [41]. A classifier infers gender from video, images, or text, and the process is usually straightforward. First, data such as video or images are parsed into a GCS. Using a feature extraction algorithm, it then extracts features from the data, such as static body features, dynamic body features, apparel features, and biometrics [36, 38, 39]. Finally, it compares those features using a classifier to a feature dataset, which is categorized by gender, and maps them to either category, inferring gender based on similarities in features [34, 51].

Similarly, a text-based GCS infers gender using features such as language, vocabulary, and frequency of words [39]. Text-based GCSs extract features using text mining from content found in forums, chat rooms, and social media [39, 50]. Beyond language, Corney *et al.* extended text-based feature extraction further into the typography field, training a classifier to make gender inferences based on style markers, structural characteristics, and gender-preferential language [12].

In the literature, developers have used classifiers to support text analysis techniques (e.g., sentiment and content analysis). Park *et al.* developed a GCS that supports sentiment analysis to identify the gender of persons making posts found on an online AIDS-related bulletin board [50]. The authors' GCS used a feature dataset that paired gender with the frequency of sentiment-driven words. During training, the GCS learned that women tended to use the words "thank," "bless," "scary," and, "illness" about twice as often as men, who themselves used "accurate," "important," "issue," and "aches" twice as often as women [50].

Several studies have made use of freely available Twitter user posts (or tweets) to train a GCS and infer the gender of other users [17, 19, 40, 45]. Lopes Filho *et al.* utilized a dataset categorizing gender by 60 textual meta-attributes associated with characters, syntax, words, structure, and morphology for the extraction of gender expression linguistic cues in tweets [40]. The authors compared different classifiers, finding that each accurately determined the gender of Twitter users, 63.5%, 61.96%, and 68.08% of the time. Using word unigrams, hashtags, and psychometric properties as features, the GCS developed by Fink *et al.* predicted the gender of Twitter users with 80% accuracy [17].

Gender recognition can be useful to support applications, such as face recognition and smart human-computer interface aid in other domains [51]. Developers use algorithmic gender classification in human-computer interaction, the security and surveillance industry, law enforcement, psychiatry, demographic research, education, commercial development, telecommunication, and mobile application and video games [34, 39, 51]. Depending on the application and dataset, developers may also use vision-based and biological information-based methods to make inferences [39].

However, "sex," "gender," and "sexuality" are often confused and used in overlapping ways, both by laypeople and experts. In this paper, we draw on the following definitions: "sex" usually refers to the assigned gender at birth based on medical factors (e.g. genitalia, chromosomes, and hormones), usually "m[20]ale" or "female" although in some cases "intersex." Sex can also be changed through medical intervention. "Gender" is both a "person's internal held sense of their gender"—also called gender identity—but is also tied to social, cultural and legal factors. "Sexuality" we take to mean the "physical, romantic, and/or emotional attraction to another person" [60]. We take into account that these definitions are also socially constructed through societal demands and norms.

3. METHODS

Available scientific literature focuses on how gender can be inferred from user attributes [17, 19, 40, 45]. However, there are not many studies that have compared the users' reported gender, the inferred gender from those attributes, and its correctness, although this avenue of research has been of an increasing interest in social sciences [23]. How algorithms exacerbate existing biases and affect marginalized communities is also a nascent area of specialization [23, 46, 64]. Our work contributes to the literature on algorithmic bias and discrimination by exploring misgendering on social media platforms like Twitter.

Given that gender stereotypes persist online, and that the social media platform Twitter infers gender from a wide variety of sources, we wondered how accurate Twitters' gender inferences of its users' gender identities are and, with the support of survey data, we explore what implications this social media practice has—such as the reinforcement of gender binarism and exacerbation of gender stereotypes [23]. We also refer to privacy and discrimination law, focusing on the impact of online behavioral advertising on inadvertent privacy violations [63] and the reinforcement of social prejudices.

We conducted a short survey disseminated using Twitter. For four days, from 22 to 26 May 2020, N=109 Twitter users responded. The online survey was prepared in Qualtrics and included five specific questions revolving around whether Twitter algorithms inferred users' gender and whether it was correct. In particular, we asked the user's sexual orientation (Q1), their gender identity (Q2), the pronouns they use (Q3), whether they provided Twitter with their gender information (Q4), and, if not, whether that was correctly assigned (Q5).[2] We gave the users instructions on how to find their assigned gender on Twitter,[3] and we processed anonymous data and surveyed the adult population only.

At the end of the survey, we exported, tabulated, and analyzed the data using Microsoft Excel Spreadsheet Software. The lead author analyzed the survey data, and the remaining authors examined the tabulated data and analysis to discuss discrepancies and ensure the reliability of the results. Empirically ascertaining if Twitter (mis)genders users lays the foundation for future work into the potential impacts in an extensive survey.

All of the respondents completed the survey in its entirety. However, the online survey has some limitations, including the small number of the respondents, which only amounts to N=109. This is due to the quick nature of our survey, within a limited, four-day timeframe. Another limitation may be the limited representativeness of the sample, which seems to over-represent the LGBTQIA+ community compared to the number of straight people in society in general. This potential bias may be due to one or more of the following reasons. First, the LGBTQIA+ community may be overrepresented among Twitter users (Twitter does not provide data on this) or be overrepresented in the authors' Twitter networks. Moreover, people from the LGBTQIA+ community may be more inclined to complete the survey, perhaps because the survey topic

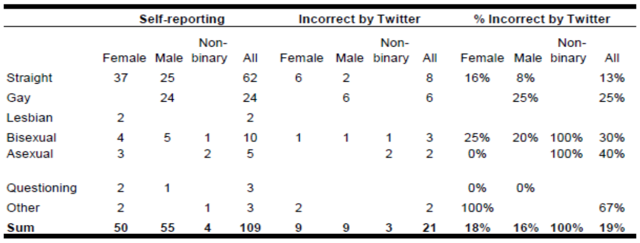
[2] The exact wording of the questions were: 1) What is your sexual orientation?; 2) What is your gender identity?; 3) What pronouns do you use?; 4) Did you at one point provide Twitter with your gender?; 5) If you did not include your gender in your profile, the gender that appears in your profile may have been assigned by Twitter, is the gender appearing here correct?

[3] To know the gender assigned by Twitter, go to picture > settings and privacy > account > your Twitter data > password > account > confirm password > gender.

appealed to them, as it may relate to past experiences of gender stereotyping or misrepresentation, on Twitter or elsewhere.

4. RESULTS

Based on the conducted online survey, we identify the following data:



	Self-reporting				Incorrect by Twitter				% Incorrect by Twitter			
	Female	Male	Non-binary	All	Female	Male	Non-binary	All	Female	Male	Non-binary	All
Straight	37	25		62	6	2		8	16%	8%		13%
Gay		24		24		6		6		25%		25%
Lesbian	2			2								
Bisexual	4	5	1	10	1	1	1	3	25%	20%	100%	30%
Asexual	3		2	5			2	2	0%		100%	40%
Questioning	2	1		3					0%	0%		
Other	2		1	3	2			2	100%			67%
Sum	50	55	4	109	9	9	3	21	18%	16%	100%	19%

Table 1. Twitter gender inference accuracy in a N=109 sample data.

Out of N=109 respondents, 19% had their gender wrongly assigned, whereas Twitter inferred users' gender correctly in 81% of the cases. Our central hypothesis revolved around differences between the self-reported gender identity (male), the sexual orientation of users (gay), and the correctness of the Twitter assigned gender (female).

Twitter infers their users' identity from a wide variety of sources, such as information from the account, interactions with links, and cookie data,[4] but not from their sexual orientation. However, how apparently fair algorithmic designs and categorizations have ulterior and unintended consequences in specific communities is well-known in the literature [10, 21, 28]. For instance, our collected data shows that, out of the misgendered Twitter users that we analyzed, only 38% were straight. Only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% straight women. Individuals that self-reported as bisexuals were misgendered in 25% of the cases for bisexual women and 20% for bisexual men. Respondents identifying as non-binary were misgendered in all cases. These results show that the LGBTQIA+ community and straight women were more often misgendered than straight men in our sample. Therefore, misgendering was, contrary to our hypothesis, not only limited to gay men compared to straight men. Moreover, women and non-binary are usually more misgendered by Twitter than men.

The findings also seem to suggest that lesbian and questioning people are less likely to be misgendered, although the numbers are small in our sample (two lesbian and three questioning participants)—more studies are needed for this sample population. One questioning and one lesbian participant answered that they had provided their gender to Twitter, meaning the gender of the remaining were inferred correctly by Twitter. The findings also show that non-binary participants (N=2) were misgendered, both of whom were also asexual participants. However, there were other asexual participants (N=3) whose gender (female in all cases) was correctly inferred by Twitter (each answered 'I do not know' when asked if they provided Twitter with their gender).

Of the 109 participants, only 15% provided Twitter with their gender, whereas 24% did not, and 61% did not remember doing so. 42% of those who did not provide Twitter with gender were from the LGBTQAI+ community. Of the 16 participants who provided their gender, all but one answered "Yes" about whether or not the gender appearing on their Twitter profile was correct. An outlier was an asexual, nonbinary person. This may indicate that either (1) some of those 16 participants were mistaken and had entered their gender into their Twitter profile previously or (2) Twitter may infer gender and change the one entered by the user.

Other findings resulted from discussions over Twitter, where we shared the online survey. Some respondents openly reported that Twitter used to misgender them, but that now Twitter gendered them correctly, probably due to their increasing interest in gender equality. Other respondents mentioned they had two profiles, but that Twitter misgendered the profile they used the most. A respondent suggested that, although gay, Twitter assigned his gender correctly, while another was surprised to be considered "female" while being a "male."

5. DISCUSSION

5.1 Misgendering in social media is discriminatory

Research affirms that gender identity is primarily subjective and internal, which juxtaposes with the idea that gender can be recognized automatically, at least with the state of art GCS [23]. Moreover, misgendering users via automated gender recognition systems have adverse implications, some of those being that they reinforce gender binarism, undermine autonomy, are a tool for surveillance, and threaten safety [23].

They also exacerbate existing stereotypes. Classifiers trained on real-world datasets are often biased because the data used to train them contains racial and gender stereotypes [7, 18, 43, 57]. Female names are more associated with family than career words, with arts more so than mathematics and science [47, 48]. Datasets imSitu and MS-COCO are significantly gender-biased and "models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data" [65]. For example, the verb "cooking" is heavily biased towards women in a classifier trained using the imSitu dataset, amplifying existing gender stereotypes [65]. The same gender biases have been shown in natural language processing [55, 66], another method used to support gender classifiers [11].

To be misgendered reinforces also the idea that society does not consider or recognize a person's gender as "real," causing rejection, impacting self-esteem and confidence, felt authenticity, and increasing one's perception of being socially stigmatized [33]. If not addressed carefully, these gender biases in the offline world may propagate to artificial intelligence [10]. This is especially concerning given that available research suggests that many individuals perceive automatic misgendering as more harmful than human misgendering [23].

Moreover, when the tools used to extract patterns and profiles from data are not transparent, it may be hard for people to contest any decisions resulting from this, which may impede their freedom and autonomy and may inadvertently affect their privacy. In the EU, the collecting and processing of personal data are protected under the General Data Protection Regulation (GDPR), which also addresses discrimination issues in datasets. However, enforcing legislation in

[4] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases

such cases is very challenging. For data protection, scholars note that information about a person's gender, age, financial situation, geolocation, and online profiles are not sensitive data according to Article 9 of the GDPR, despite often being grounds for discrimination [63]. Not being "sensitive data" translates into not enjoying the extra protection (such as users' informed and explicit consent) that categories of information deemed sensitive such as race, religion, or sexual orientation have. Discrimination in (patterns and profiles extracted from) large datasets can be hard to detect. Indirect discrimination takes place unintentionally when users are unaware of any harm profiles may be doing. However, it may also be the case that companies use profiles precisely to conceal discrimination, a process called masking [13]. Because direct discrimination in data is hard to detect, and indirect discrimination is nearly impossible to detect, it can be challenging to enforce equal treatment acts and data protection legislation.

Many forms of discrimination are illegal in most Western jurisdictions. Not hiring someone based on their gender, ethnicity, or sexual orientation, or because they have a criminal record, is prohibited for most professions. Not every decision based on the sensitive characteristics mentioned is forbidden, however. Legislation that forbids discrimination based specific characteristics lists the characteristics that may not serve as a basis for making decisions (including gender, ethnicity, political preferences, trade union membership, or sexual orientation). Nonetheless, "softer" forms of discrimination, in the form of stigmatization of specific population groups may occur, for example, in the formation of friendships. On a larger scale, this could lead to social polarization and segregation. For now, misgendering or addressing someone with the wrong pronoun is not sufficiently grave to be considered harassment under certain specific legal provisions (although there has been advocacyfor remedies for these acts [4]).

5.2 Inferences Organizations controversially infer gender for legitimate interests

Twitter makes inferences about users' accounts, including interests, age, and gender, to provide features such as account suggestions (e.g., suggested contacts, promoted accounts for the user to follow), advertising, recommendations, and timeline ranking.[5] Twitter uses users' content, activity, relationships, and interactions to genderize content production patterns [52], infer gender, and make these suggestions.[6] Twitter justifies making inferences about interests, age, and gender, stating that it helps tailor content to users, keeps the platform safe and enjoyable for all users, and enables Twitter to provide compelling, targeted advertising. In other words, users have to accept the trade-off if they want to have a personalized Twitter account.

The GDPR lists a limited number of legal grounds for data processing, including consent, the performance of a contract, or legitimate interests. Twitter states that it makes "inferences about your account - such as interests, age, and gender" for "legitimate purposes." The appeal to legitimate interests as a legal basis for data processing is controversial, as legitimate purposes are only a solid legal basis if there is a necessity. It is questionable, however, whether gender inferences are necessary for Twitter. Although the legitimate interest seems less constraining than other grounds for data processing, it should not be considered a "last resort" when all other grounds for lawful data processing fail [3].

Legitimate interest is the most appropriate legal ground for data processing if the data controller uses people's data in ways they would reasonably expect and have a minimal privacy impact, or where there is a compelling justification for the processing. If controllers choose this legal ground, they "should take on extra responsibility for considering and protecting people's rights and interests" [27]. Thus, three elements configure the basis for legitimate interest: identifying the legitimate interest, showing that the processing is necessary to achieve it, and balancing it against the individual's interests, rights, and freedoms.

Our survey findings highlight a significant number of misgendered users and question whether Twitter did balance their interests against individuals' interests. First, out of the 109 participants, only 15% provided Twitter with their gender, while Twitter inferred their gender anyway. Second, our results suggest that the LGBTQIA+ community and straight women may be more often misgendered than straight men. Third, remedies for opposing the processing seem not to correspond in magnitude to the subsequent impact of being misgendered. A user can modify or rectify the inferred gender but cannot escape that inference unless she actively opts out of Twitter's personalization features. Making users choose between these two is as if, in times of COVID-19, developers made users choose between health or privacy [25]. Moreover, it results in a privacy paradox: the gender inference causes a privacy issue (i.e., disclosing information people may want to keep to themselves), but to address this, users have to provide additional information, disclosing even more (or more detailed) information about themselves [13]. This is particularly problematic for communities that society has been historically discriminated against and in which gender is a sensitive part of their identity [16, 44].

5.3 Accounting for diversity in social media

Platforms exclude and misrepresent a large number of potential users if they are not respectful and inclusive towards their gender identity or sexual orientation. The assumption that gender is physiologically-rooted harms trans people overall by essentializing the body as the source of gender, and also harms non-binary people, who cannot be accurately classified [33]. As Fergus highlighted, transgender and non-binary users reported being misgendered by Twitter, which we found to be the case in our survey (100% of the non-binary participants reported being misgendered) [16]. These findings may result from the fact that Twitter gender classifiers do not account for diversity and work on male/female binary categorization that, although it does represent some people's gender expression, does not do justice to the freedom of identity that everyone should have.

Our study shows that when it comes to diversity and more inclusive engagement, social media platforms like Twitter still have a long way to go to become a more open and welcoming platform for a wide variety of users. Misgendering users in the background is not good practice, and beyond echoing deeply rooted stereotypes, can lead to privacy and discrimination issues,. The lack of diversity in marketing strategies is apparent when users can be gendered as male or female only. However, making strategies for a diverse

[5] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases; *see* also https://help.twitter.com/en/using-twitter/account-suggestions.

[6] *See* https://help.twitter.com/en/using-twitter/account-suggestions

engagement with the "queer rainbow economy" can make for more affluent and more diverse revenue streams [6, 32, 58].

From all this, it is clear that digital identity and participatory culture play a massive role in the sense of self in the modern world and that there should be more effort to realize diversity and inclusion in the online world [29] to not perpetuate the normative view that certain groups of people, such as trans or non-binary people, do not exist [33].

6. SUMMARY

An online survey showed that, out of N=109 respondents, Twitter correctly inferred users' gender in 81% of the cases, and 19% were misgendered. A close look at the results shows that only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% of straight women, while non-binary users were misgendered in all the cases.

Social media platforms like Twitter have economic incentives to know users' genders for commercialization and targeted advertisements. However, our investigation shows that inferring a user's gender with automated means clashes with the understanding that gender is subjective and internal. Misgendering has also broader consequences, leading to serious privacy, discrimination, autonomy, and self-identity issues. Misgendering reinforces gender stereotypes, accentuates gender binarism, undermines autonomy, and leads to toxic cultures and algorithmic bias [23, 37]. Moreover, misgendering causes a feeling of rejection, impacting one's self-esteem, confidence, and authenticity, increasing social stigmatization [33].

If users do not provide a gender parameter choice themselves, platforms may infer the user's gender from a wide variety of data sources, including personal data. Therefore, gender classifiers should account for diversity and inclusion, using a more accurate understanding of gender to represent contemporary society fully. Otherwise, inferential analytics may reinforce existing biases about gender stereotyping. By including diverse users early on, during the design, and with the possibility to provide feedback afterward, the technology can be experienced as more just and fairer. Inclusive engagement that reflects on the users as not homogeneous can have a positive impact on technology.

By identifying how inferential analytics may reinforce gender stereotyping and affect marginalized communities, we hope to continuously contribute to promoting the online account for privacy, diversity, and inclusion and advocate for the freedom of identity that everyone should have online and offline [69]. Looking forward, a more robust survey ought to be undertaken to further explore the social implications of gender inference on Twitter, such as discrimination and diversity in social media.

7. ACKNOWLEDGMENTS

Part of this project was funded by the LEaDing Fellows Marie Curie COFUND fellowship, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 707404.

8. REFERENCES

[1] *Convention on the elimination of all forms of discrimination against women.* https://www.ohchr.org/en/professionalinterest/pages/cedaw.aspx#:~:text=On%2018%20December%201979%2C%20the,twentieth%20country%20had%20ratified%20it, 1979.

[2] *Convention on the rights of persons with disabilities.* https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-8-awareness-raising.html, 2008.

[3] Article 29 Data Protection Working Party. *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf, 2014.

[4] Ashley, F. *No, pronouns won't send you to jail: The misunderstood scope of Bill C-16*. Medium, https://medium.com/@florence.ashley/no-pronouns-wont-send-you-to-jail-43c268cffd55, 2017.

[5] Bray, F. Gender and technology. *Annual Review of Anthropology*, 36(1): 37-53, 2007.

[6] Brown, G. Thinking beyond homonormativity: Performative explorations of diverse gay economies. *Environment and Planning A: Economy and Space*, 41(6): 1496-1510, 2009.

[7] Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the First Conference on Fairness, Accountability and Transparency*, 77–91. PMLR, 2018.

[8] Butler, J. *Gender trouble, feminist theory, and psychoanalytic discourse*. Routledge, 1990.

[9] Calders, T. and Custers, B. *What Is data mining and how does it work?* In: *Discrimination and privacy in the information society: Data mining and profiling in large databases* (eds. Custers et al.), pages 27–42. Springer, 2013.

[10] Caliskan, A., Bryson, J. J. and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183-186, 2017.

[11] Campa, S., Davis, M. and Gonzalez, D. *Deep & machine learning approaches to analyzing gender representations in journalism*.https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf, 2019.

[12] Corney, M., Vel, O. d., Anderson, A. and Mohay, G. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference, 2002*, pages 282-289, 2002.

[13] Custers, B. Data dilemmas in the information society: Introduction and overview. In B. Custers, T. Calders, B. Schermer and T. Zarsky, editors, *Discrimination and privacy in the information society*, pages 3–26, Springer, 2013.

[14] Custers, B. *Profiling as inferred data: Amplifier effects and positive feedback loops*. Amsterdam University Press, http://www.jstor.org/stable/j.ctvhrd092.23, 2018.

[15] Davenport, T. D. and Beck, J. C. *The attention economy: Understanding the new currency of business*. Harvard Business School Press, 2002.

[16] Fergus, J. *Twitter is guessing users' genders to sell ads and often getting it wrong*. Input, https://www.inputmag.com/tech/twitter-guesses-your-gender-to-serve-you-ads-relevant-tweets-wrong-misgendered, 2020.

[17] Fink, C., Kopecky, J. and Morawski, M. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Medi*, pages 459–452, AAAI, 2012.

[18] Font, J. E. and Costa-jussà, M. R. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 147–154. Association for Computational Linguistics, 2019.

[19] Garibo i Orts, O. A big data approach to gender classification in Twitter. In *Proceedings of the Ninth International Conference of the CLEF Association*, 2018.

[20] GLAAD *GLAAD media reference guide – transgender*. https://www.glaad.org/reference/transgender, n.d.

[21] Gomes, A., Antonialli, D. and Olivia, T. D. *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/, 2019.

[22] Grant, A., Grey, S. and van Hell, J. G. Male fashionistas and female football fans: Gender stereotypes affect neurophysiological correlates of semantic processing during speech comprehension. *Journal of Neurolinguistics*, 53: 100876, 2020.

[23] Hamidi, F., Scheuerman, M. K. and Branham, S. M. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper 8, ACM, 2018.

[24] Hao, K. *Facebook's ad-serving algorithm discriminates by gender and race*. MIT Technology Review, https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/, 2019.

[25] Harari, Y. N. *The world after coronavirus*. https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75, 2020.

[26] Hentschel, T., Heilman, M. E. and Peus, C. V. The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology*, 10(11), 2019.

[27] Information Commissioner's Office *Legitimate interests*. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/legitimate-interests/, n.d.

[28] Ito, J. *Supposedly 'fair' algorithms can perpetuate discrimination*. Wired, https://www.wired.com/story/ideas-joi-ito-insurance-algorithms/, 2019.

[29] Jenkins, H., Ito, M. and boyd, d. *Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics*. Polity Press, Cambridge, UK, 2016.

[30] Jernigan, C. and Mistree, B.F.T. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[31] Kamiran, F., Calders, T. and Pechenizkiy, M. Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. Schermer and T. Zarsky, editors, *Discrimination and privacy in the information society*, pages 223–239, Berlin, Springer, Berlin, Heidelberg. 2013.

[32] Keating, A. and McLoughlin, D. Understanding the emergence of markets: A social constructionist perspective on gay economy. *Consumption Markets & Culture*, 8(2): 131-152, 2005.

[33] Keyes, O. The misgendering machines: Trans/HCI Implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW): Article 88, 2018.

[34] Khan, S. A., Ahmad, M., Nazir, M. and Riaz, N. A comparative analysis of gender classification techniques. *International Journal of Bio-Science and Bio-Technology*, 5(4): 223–244, 2013.

[35] Kosinski, M., Stillwell, D. and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15): 5802-5805, 2013.

[36] Kumar, D., Gupta, R., Sharma, A. and Saroj, S. K. Gender classification using skin patterns. In *Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019.

[37] Lambrecht, A. and Tucker, C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7): 2966-2981, 2019.

[38] Li, B., Lian, X.-C. and Lu, B.-L. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 76(1): 18-27, 2012.

[39] Lin, F., Wu, Y., Zhuang, Y., Long, X. and Xu, W. Human gender classification: A review. *International Journal of Biometrics*, 8(3-4): 275–300, 2016.

[40] Lopes Filho, J. A. B., Pasti, R. and de Castro, L. N. Gender classification of Twitter data based on textual meta-attributes extraction. In Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis and M. Mendonça Teixeira, editors, *New advances in information systems and technologies*, pages 1025–1034, Cham, Springer International Publishing, 2016.

[41] Mathivanan, P. and Poornima, K. Biometric authentication for gender classification techniques: A review. *Journal of The Institution of Engineers (India): Series B*, 99(1): 79-85, 2018.

[42] Matz, S. C., Menges, J. I., Stillwell, D. J. and Schwartz, H. A. Predicting individual-level income from Facebook profiles. *PLOS ONE*, 14(3): e0214369, 2019.

[43] McDuff, D., Song, Y., Kapoor, A. and Ma, S. Characterizing bias in classifiers using generative models. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

[44] McLemore, K. A. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1): 51-74, 2015.

[45] Nieuwenhuis, M. and Wilkens, J. Twitter text and image gender classification with a logistic regression n-gram model. In *Proceedings of the Ninth International Conference of the CLEF Association*, pages, 2018.

[46] Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, New York, 2018.

[47] Nosek, B. A., Banaji, M. R. and Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1): 101-115, 2002.

[48] Nosek, B. A., Banaji, M. R. and Greenwald, A. G. Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, 83(1): 44-59, 2002.

[49] O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, New York, 2016.

[50] Park, S. and Woo, J. Gender classification using sentiment analysis and deep learning in a health web forum. *Applied Sciences*, 9, 2019.

[51] Rai, P. and Khanna, P. Gender classification techniques: A review. In D.C. Wyld et al. (editors), *Advances in computer science, engineering & applications*, pages 51–59, Berlin, Springer, 2012.

[52] Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M. and Stern, M. J. Digital inequalities and why they matter. *Information, Communication & Society*, 18(5): 569-582, 2015.

[53] Roosendaal, A. Facebook tracks and traces everyone: Like this! *Tilburg Law School Legal Studies Research Paper Series*, 2010.

[54] Schiebinger, L. Scientific research must take gender into account. *Nature*, 507(7490): 9, 2014.

[55] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. and Wang, W. Y. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, 2019.

[56] Thorson, K., Cotter, K., Medeiros, M. and Pak, C. Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*: 1-18, 2019.

[57] Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pages 1521–1528, 2011.

[58] Um, N.-H. Seeking the holy grail through gay and lesbian consumers: An exploratory content analysis of ads with gay/lesbian-specific content. *Journal of Marketing Communications*, 18(2): 133-149, 2012.

[59] UN Office of the High Commissioner Human Rights *Gender stereotyping*. https://www.ohchr.org/EN/Issues/Women/WRGS/Pages/GenderStereotypes.aspx, n.d.

[60] University of Washington Human Resources Office *Terminology*. https://hr.uw.edu/ops/transgender-resources/terminology/, n.d.

[61] Ur, B., Leon, P. G., Cranor, L. F., Shay, R. and Wang, Y. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, Article 4. ACM, 2012.

[62] Wachter, S. Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal*, 35(2), 2020.

[63] Wachter, S. and Mittelstadt, B. A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2): 494–620, 2019.

[64] Willson, M. Algorithms (and the) everyday. *Information, Communication & Society*, 20(1): 137-150, 2017.

[65] Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K. W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

[66] Zhou, P., Zhao, J., Huang, K.-H. and Shi, W. Examining gender bias in languages with grammatical gender. 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 5279-5287, 2019.

[67] Žliobaitė, I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2): 183-201, 2016.

[68] Zuboff, S. Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1): 75-89, 2015.

[69] Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., & Custers, B. H. M. (2021). A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, *58*(3), 102541.

About the authors:

Dr. Eduard Fosch-Villaronga is an Assistant Professor at the eLaw Center for Law and Digital Technologies at Leiden University (NL). Fosch-Villaronga investigates the legal and regulatory aspects of robot and Artificial Intelligence (AI) technologies and is the leader of the Robotics and Autonomous Systems Working Group, which is an interdisciplinary Working Group at eLaw dedicated to advance the understanding of the legal, regulatory, and ethical implications of robots and autonomous systems. Previously, he was a Marie Sklodowska-Curie postdoctoral researcher at the same group and served the European Commission in the Sub-Group on Artificial Intelligence (AI), connected products and other new challenges in product safety to the Consumer Safety Network (CSN) to revise the General Product Safety directive. He is also the co-leader of the Working Group on the Ethical, Legal and Societal Aspects for Wearable Robots at the H2020 COST Action CA16116. Among his publications, Fosch-Villaronga published a book entitled *Robots, Healthcare, and the Law: Regulating Automation in Personal Care* with Routledge.

Mr. Adam Poulsen is a computer scientist & PhD candidate at Charles Sturt University, Australia. Poulsen's research focus is on value sensitive robots, LGBTQIA+ elders and care, and robot and machine ethics. Poulsen is a recipient of the Australian Government Research Training Program Scholarship. Among other initiatives, Poulsen has contributed actively to the Australian Association of Gerontology's Assistive Technology Special Interest Group as a co-convenor.

Dr. Roger A. Søraa is a researcher at the Department of Interdisciplinary Studies of Culture, NTNU Norwegian University of Science and Technology. He also works as a project leader and project developer for the Department of Neuromedicine and Movement Science at the same university. An ongoing collaboration between these departments is the Immersive Technologies and Robotics laboratory (ImRo), where Søraa serves as the deputy manager. Next to his positions at NTNU, Søraa is affiliated with RURALIS, where he researches smart technology for sustainable agriculture. Søraa completed his PhD dissertation in Studies of Science and Technology in 2018. Søraa's current research interests include robotization of gerontechnologies-, transport and agriculture, automation of work, and practices and digitalization of society and social media.

Prof. Dr. Bart H. M. Custers is a full professor of law and data science and director of eLaw, the Center for Law and Digital Technologies at Leiden University, the Netherlands. With a background in both law and physics, his research is focused on law and digital technologies. Research interests include discrimination and privacy issues of new technologies, particularly data mining and profiling. Dr. Custers has published seven books, including four books on discrimination and privacy in the context of Big Data. On a regular basis, he gives lectures on profiling and privacy issues of new technological developments. He presented his work at international conferences in the United States, Canada, China, Japan, the Middle East, Africa, and throughout Europe. He has published his work, over 100 publications, in both scientific and professional journals and newspapers.