eLaw
Working Paper Series

No 2021/00 ! - ELAW– 2021

M,xu¡¡xq,nu~p,¡{xp, y q,¦{¢~,sqzpq~
Sqzpq~,uzrq~qzoq ,uz, {oumx, y qpum
R{ ot9buxxm~{zsm8,Q:8,\{¢x qz8,M:8,_Ě~mm8,^:M:8,
O¢ ¡q~ 8,N:T:Y:,

Universiteit Leiden
eLaw

Discover the world at Leiden University

Information Processing and Management 58 (2021) 102541

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A little bird told me your gender: Gender inferences in social media

E. Fosch-Villaronga [a,*], A. Poulsen [b], R.A. Søraa [c], B.H.M. Custers [a]

[a] eLaw Center for Law and Digital Technologies, Leiden University, the Netherlands
[b] School of Computing and Mathematics, Charles Sturt University, Australia
[c] Researcher at Department of Interdisciplinary Studies of Culture and Department Neuromedicine and Movement Science, Norwegian University of Science and Technology (NTNU), Norway

ARTICLE INFO

Keywords:
Gender
Twitter
Social media
Inference
Gender classifier
Automated gender recognition system
Privacy
Algorithmic bias
Discrimination
LGBTQAI+
Gender stereotyping
Online Behavioral Advertising

ABSTRACT

Online and social media platforms employ automated recognition methods to presume user preferences, sensitive attributes such as race, gender, sexual orientation, and opinions. These opaque methods can predict behaviors for marketing purposes and influence behavior for profit, serving attention economics but also reinforcing existing biases such as gender stereotyping. Although two international human rights treaties include explicit obligations relating to harmful and wrongful stereotyping, these stereotypes persist online and offline. By identifying how inferential analytics may reinforce gender stereotyping and affect marginalized communities, opportunities for addressing these concerns and thereby increasing privacy, diversity, and inclusion online can be explored. This is important because misgendering reinforces gender stereotypes, accentuates gender binarism, undermines privacy and autonomy, and may cause feelings of rejection, impacting people's self-esteem, confidence, and authenticity. In turn, this may increase social stigmatization. This study brings into view concerns of discrimination and exacerbation of existing biases that online platforms continue to replicate and that literature starts to highlight. The implications of misgendering on Twitter are investigated to illustrate the impact of algorithmic bias on inadvertent privacy violations and reinforcement of social prejudices of gender through a multidisciplinary perspective, including legal, computer science, and critical feminist media-studies viewpoints. An online pilot survey was conducted to better understand how accurate Twitter's gender inferences of its users' gender identities are. This served as a basis for exploring the implications of this social media practice.

1. Introduction

Online and social media platform providers use attributes of their users, including their name, age, and gender, to improve user experience and online behavioral advertising (OBA). For instance, the social media platform Twitter infers gender from a wide variety of sources.[1] By processing user attributes, companies can target or exclude certain groups more easily, tailor their services to users, and increase their attention levels (Ur, Leon, Cranor, Shay & Wang, 2012). In this way, profiling makes marketing more precise and

* Corresponding author.
E-mail addresses: e.fosch.villaronga@law.leidenuniv.nl (E. Fosch-Villaronga), apoulsen@csu.edu.au (A. Poulsen), roger.soraa@ntnu.no (R.A. Søraa), b.h.m.custers@law.leidenuniv.nl (B.H.M. Custers).

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

https://doi.org/10.1016/j.ipm.2021.102541
Received 29 October 2020; Received in revised form 7 January 2021; Accepted 2 February 2021
Available online 18 February 2021

effective. However, a growing concern is the increasing use of *transparent* inferential analytics that reveal sensitive user traits that serve attention economics,[2] (Davenport & Beck, 2001) and that may reinforce existing biases that, although not explicit, can be very influential in exacerbating discrimination (Caliskan, Bryson & Narayanan, 2017; Custers, 2018).

One bias is gender stereotyping, which 'refers to the practice of ascribing to an individual woman or man specific attributes, characteristics, or roles by reason only of her or his membership in the social group of women or men' (OHCHR, 2020). However, gender stereotyping is a complex process that, although based on strong beliefs of what gender is and should be, is understood too simplistically (Kachel, Steffens & Niedlich, 2016). For instance, Sink, Mastro and Dragojevic (2018): 592) investigated how television character perceptions often judged effeminate gay men negatively. They found out that 'straight-acting' or hyper-masculine gay men are evaluated more favorably for conforming to and even mastering heteronormative gender roles. Also, "gay men who are perceived to be more feminine would map onto traditional female stereotypes (i.e., warm but less competent)" (ibid).

Two international human rights treaties include explicit obligations relating to harmful and wrongful stereotyping (mainly Art. 5 of the Convention on the Elimination of All Forms of Discrimination against Women and Art. 8(1)(b) of the Convention on the Rights of Persons with Disabilities).[3] Although States are usually the recipients of human rights treaties, the United Nations Human Rights Council has shown growing attention to the responsibility that corporations, sectors, and industries worldwide have for respecting human rights (OHCHR, 2012). Still, these stereotypes persist online and offline (Grant, Grey & van Hell, 2020; Hentschel, Heilman & Peus, 2019), as if platforms failed to understand—or deliberately choose to ignore—that gender is not merely being a 'man' or a 'woman,' but a social construct (Butler, 1990). In the words of De Beauvoir (1949), "one is not born, but rather becomes a woman". It is now becoming clear that these practices continue to exist online, for instance, on social media platforms. As described below, Twitter often assigns you to be a 'woman' if you're a gay man. This socio-technical construction of gender shows the intricate gendered way of how humans shape their own identities through the technology they use and interact with (Søraa, 2017).

In this contribution, the impact of inferential analytics on inadvertent privacy violations and the reinforcement of social prejudices of gender and sexuality is investigated by means of a specific case study (i.e., Twitter) through a multidisciplinary perspective, including legal, computer science, and queer media viewpoint. Therefore, the central research question of this contribution is: *What are the implications of any inaccurate gender inferences by Twitter?* This research question is only relevant if Twitter actually performs gender inferences and if these inferences are sometimes inaccurate. As is shown in a pilot survey presented in this contribution, there is evidence supporting this. In a broader perspective, addressing this research question illustrates concerns of discrimination, misgendering, and exacerbation of existing biases that online platforms persist in replicating and that literature has started to highlight very recently (Hamidi, Scheuerman & Branham, 2018; Keyes, 2018).

The rationale behind the research question is, first, 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. The assumption that gender is physiologically-rooted harms transgender people by essentializing the body as the source of gender and also harms non-binary people, who cannot be accurately classified (Keyes, 2018). The categories 'female' and 'male' do not reflect who they are (Fergus, 2020). Second, that platform providers no longer have to learn sensitive details about a particular user or to correctly group users into categories for advertising to be effective, as advertising has a high tolerance for classification errors (Wachter, 2020). Nevertheless, not considering gender and sexuality in social media platforms can be socially harmful and expensive, as the unconscious bias of technology usage and implementation may lead to further exacerbation of existing biases, for gender and race if only (Bray, 2007; Hao, 2019a; Schiebinger, 2014).

In this article, we start with providing some background information on the mechanics of inferential analytics to elucidate how companies infer specific user attributes, including gender, and how these techniques may harm users' rights in Section 2. Also, we introduce the technical rationale of gender classifiers. In Section 3, we explain how we conducted this study, which is based on configuring an online pilot survey geared towards understanding how accurate Twitters' gender inferences of its users' gender identities are and, with the support of survey data, explore what implications this social media practice has. Our results, presented in Section 4, already anticipate the binary understanding of gender from Twitter, which excludes those not fitting the category 'male' and 'female,' that inferring gender is part of the Twitter's personalization trade-off, and that in nearly 20% of the cases it misgendered its users. After presenting the data, we discuss the social implications of gender inference on social media platforms in Section 5, including the lack of diversity in social media platforms and the role designers play in accounting for inclusivity and diversity. We also argue that platforms use a form of scapegoating to get away with the inference of sensitive user traits without user awareness. Consequences for data protection and discrimination are also discussed. We conclude the article in Section 6 by summarizing the most important takeaways from this article and presenting our future work.

2. Gendering algorithms

Organizations worldwide employ inferential analytics methods to guess user characteristics and preferences, including sensitive

[2] According to Lexico, in computing, *transparent* means "(of a process or interface) functioning without the user being aware of its presence." In this article, we will use the word *opaque* to stress the user's unawareness. *See* https://www.lexico.com/definition/transparent.

[3] Art. 5 of the Convention on the Elimination of All Forms of Discrimination states 'States Parties shall take all appropriate measures: (a) To modify the social and cultural patterns of conduct of men and women, with a view to achieving 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 on stereotyped roles for men and women;' and Art. 8(1)(b) 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.'

attributes, such as race, gender, sexual orientation, and political opinions. Questions regarding the consequences of automated gender recognition are particularly poorly understood and often underestimated because technical approaches often miss a broader social perspective, especially concerning sex and gender (Nature, 2020; Tannenbaum, Ellis, Eyssel, Zou & Schiebinger, 2019). At the same time, the global landscape of AI ethics guidelines does not provide adequate guidance in this respect (Jobin, Ienca & Vayena, 2019), partly because the social science community and the policymaking do not always understand the inner workings of these practices and its ulterior implications, which tend to be black-boxed.

Understanding how the ever-increasing use of inferential analytics supporting algorithmic decision-making processes may reinforce gender stereotyping and affect marginalized communities worldwide requires, therefore, a multidisciplinary perspective. This section brings together three different perspectives and compiles useful information to help further understand this issue. First, we explain briefly what gender and sex are from a queer media-studies viewpoint (Section 2.1). Next, we explain what automated gender recognition is and how gender classifiers work from a computer science perspective (Section 2.2). Finally, we explain how organizations and companies use these methods to support ulterior decision-making processes that significantly affect citizens in various ways from a more social science outlook (Section 2.3).

2.1. Understanding sex and gender

Scientific research is increasingly taking gender and sex into account because it makes better science (Schiebinger, 2014; Tannenbaum et al., 2019). However, queer media-studies research stresses that "*sex*", "*gender*" and "*sexuality*" are often confused and used in overlapping ways by both 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 'male' or 'female.'—and in some cases 'intersex'. Sex is changeable through medical gender transition.
- "*Gender*" is both a "person's internal, deeply held sense of their gender," also called *gender identity*.[4]—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.'[5] However, we take into account that these definitions are also socially constructed through societal demands and norms.

Modern society also challenges the interplay between these definitions. In the intricate relations between sex, gender, and sexuality,[6] there are multiplicities of understanding, accepting, legalizing, and including diverse societal groups by both the nation-states and political levels (Hooper, 2001; Klein, 2013; Randall & Waylen, 2012), large company and corporations (Ge, Knittel, MacKenzie & Zoepf, 2016; Haas & Hwang, 2007; Ogasawara, 1998; Rosa & Dawson, 2006). This is also true for social media, digital platforms and how gender is *utilized* in and through algorithms, of which automated gender recognition systems is one example.

2.2. Automated gender recognition systems

From a computer science perspective, automated gender recognition systems usually take *sex* as a basic point of reference and have been used to boost applications, such as face recognition and smart human-computer interface (Rai & Khanna, 2012). Areas where algorithmic gender classification is being applied include human-computer interaction, the security and surveillance industry, law enforcement, psychiatry, demographic research, education, commercial development, telecommunication, and mobile application and video games (Khan, Ahmad, Nazir & Riaz, 2014; Lin, Wu, Zhuang, Long & Xu, 2015; Rai & Khanna, 2012). Depending on the application and dataset, vision-based methods and biological information-based methods might be used to make inferences (Lin et al., 2015).

Gender classification systems (GCS) are trained using a training dataset (or corpus) of structured and labelled data. These labels categorize data, and the features within, as either masculine or feminine (Rai & Khanna, 2012). Training a GCS builds a classification algorithm (or classifier) which categorizes features, such as body movements, physiological and behavioral characteristics, and facial features (Rai & Khanna, 2012), found in new data by comparing it to labelled features in the dataset. Traditionally, to make an inference, a GCS uses a feature extraction algorithm, classifier, and dataset (Lin et al., 2015). Classifiers are trained machine learning models, exemplary models include neural networks (Rai & Khanna, 2012), support vector machine (Li, Lian & Lu, 2012), K-nearest neighbor (Khan et al., 2014), and Adaboost (Mathivanan & Poornima, 2018).

A classifier infers gender from video, images, or text, and the process is usually straightforward. First, data such as video or images is parsed into a GCS. Then, using a feature extraction algorithm, it extracts features from the data, such as static body features (e.g., face, eyebrows, hand shape, body shape, skin patterns, hair), dynamic body features (e.g., gait and gesture), apparel features (e.g., clothing and footwear) and biometrics (e.g., iris, fingerprint, voice, and emotional speech) (Kumar, Gupta, Sharma & Kumar Saroj, 2019; Li et al., 2012; Lin et al., 2015). Finally, it compares those features using a classifier to a feature dataset which are categorized by gender and maps them to either category, inferring gender based on similarities in features (Khan et al., 2014; Rai & Khanna, 2012).

Similarly, text-based GCS infer gender using features such as language, vocabulary, and frequency of words (Lin et al., 2015).

[4] *See* https://www.glaad.org/reference/transgender.

[5] *Ibidem.*

[6] *See* the conceptualization of the "genderbread person" for a clear understanding of these concepts at https://www.genderbread.org/.

Text-based GCS make use of content found in emails, blogs, forums, chat rooms, and social media, extracting features using text mining (Lin et al., 2015; Park et al., 2019). Beyond language, Corney, De Vel, Anderson and Mohay (2002) 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.

In the literature, classifiers have been used to support text analysis techniques (e.g., sentiment and content analysis), which infers a gender. For instance, Park et al. (2019) developed a GCS that supports sentiment analysis to identify the gender of persons making posts found on an online AIDS-related bulletin board. The author's GCS identified gender using naïve Bayes, support vector machine, random forest, and convolution neural network classifiers, along with a feature dataset that paired gender with the frequency of sentiment-driven words. During training, the author's GCS learned that women tended to use the words 'thank,' 'bless,' 'scary,' and 'illness' about twice as often as men who used 'accurate,' 'important,' 'issue,' and 'aches' twice as often as women. As another example, Yan et al. (2006) used a naïve Bayes classifier and the frequency of words, word fonts and cases, punctuation marks, and emoticons found on blogs to identify the gender of the Internet bloggers.

Several studies in the field of computer science have made use of freely available Twitter user posts (or 'tweets') to train a GCS and thereafter infer the gender of other users (Nieuwenhuis & Wilkens, 2018; Orts, 2018; Filho, Ahirton, Pasti & De Castro, 2016; Fink, Kopecky & Morawski, 2012). Filho et al. (2016) utilized a database categorising gender by 60 textual meta-attributes associated with characters, syntax, words, structure and morphology for the extraction of gender expression linguistic cues in tweets, and thereafter gender inference. The authors compared the best first three, multinomial näive Bayes, and support vector machines classifiers, finding that each accurately determined the gender of Twitter users 63.5%, 61.96%, and 68.08% of the time, respectively. As another example, using word unigrams, hashtags, and psychometric properties as features, the support vector machine classifier developed by Fink et al. (2012) predicted the gender of Twitter users with 80% accuracy.

One of the parameters used to infer attributes from people is the 'like' button on many social media platforms (Roosendaal, 2010). In plain language, what you like tells something about who you are. Matz, Menges, Stillwell and Schwartz (2019) recently stated that Facebook likes and status updates predicted a person's income pretty accurately. Gender can also be inferred from Facebook likes with very high accuracy (Kosinski, Stillwell & Graepel, 2013). With approximately 250 Facebook likes, gender could be predicted with accuracy rates of 93%. Although these may seem many Facebook likes required, when using only five Facebook likes, gender could be predicted with accuracy rates of about 70%. Moreover, when using only one Facebook like, the accuracy rates were approximately 60% for gender predictions. According to Kosinski, Stillwell & Graepel (2013), predictions for homosexuality were about 88% accurate for gays and 75% for lesbian, and predictions on being single versus in a relationship were about 67% accurate. Given these findings' accuracy rates, predicting such sensitive attributes can result in wrong assumptions that can have ulterior consequences for users (Kosinski, Stillwell & Graepel, 2013).

2.3. *The consequences of automated recognition systems*

Our motivation for this research lies in recent social science literature that highlights that automated gender recognition systems reinforce binarism and exacerbates gender stereotyping because they use *sex* as a basis for their systems (Hamidi et al., 2018; Yekes, 2018). The binary understanding of gender on a spectrum of masculine or feminine traits reinforces gender stereotyping of different members of the LGBTQ+ community, including non-binary and trans people (Burdge, 2007; Howansky, Wilton, Young, Abrams & Clapham, 2019; Wilchek-Aviad, Tuval & Zohar, 2020), with gay men who exhibit feminine characteristics being at risk from men whose masculinity is threatened (Glick, Gangl, Gibb, Klumpner & Weinberg, 2007: 55). It can also happen that gay men are hyper-masculinized, for instance, in the "promiscuity stereotype," leading to opposition to gay rights (Pinsof & Haselton, 2017).

The implications of inaccurate gender inferences can only be identified if it is clear how the gender inferences work and how (in) accurate they actually are. However, while technical literature focuses on how gender can be inferred from user traits (Nieuwenhuis & Wilkens, 2018; Garibo-Orts, 2018; Filho et al., 2016; Fink et al., 2012), there are not many studies that juxtapose the gender that users report and the inferred gender from those traits. However, it is becoming an increasing area of interest in social sciences (Hamidi et al., 2018; Yekes, 2018). There is also little understanding of how algorithms exacerbate existing biases and affect marginalized communities, although it is a nascent area of research (Ito, 2019; Noble, 2018; Willson, 2017).

While the legal literature has not fully covered the consequences of automated gender recognition systems, it has provided a good understanding of this social media practice and the far-reaching consequences automated recognition systems or, as they call it, inferential analytics, have for society (Wachter & Mittelstadt, 2019). Companies employ automated recognition systems to infer user preferences, sensitive attributes such as race, gender, sexual orientation, political interests, and opinions (Jernigan & Mistree, 2009; Thorson, Cotter, Medeiros & Pak, 2019). Profiles and patterns extracted from large datasets are often considered useful knowledge for subsequent decision-making and micro-targeting (Hildebrandt & Gutwirth, 2008; Zarsky, 2003). These methods can predict behaviors for marketing purposes and influence behavior for profit (Wachter, 2020; Zuboff, 2015). These methods support ulterior decision-making processes that significantly affect citizens in various ways, such as the automatic refusal of an online credit application, e-recruiting practices without any human intervention, or misdiagnoses of certain diseases (Hänold, 2018).

The use of inferred data may have advantages. For instance, inferential analytics can be a tool to fill gaps in incomplete datasets or check the correctness 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 accuracy. In big data analytics, completeness and correctness of data is not a strict condition but obviously may contribute to getting more accurate 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 tailor the user experience to fit the preferences of that particular individual.

However, the use of inferred data may also bring along some issues. Privacy is typically raised as an issue when people's attributes that people did not want to disclose are predicted. Furthermore, inferred data may lead to inaccuracies, which can lead to bias and unfair decisions, and may lead to self-fulfilling prophecies, a phenomenon well-known in profiling (Custers, 2013). These effects may amplify inequality, undermine democracy, and further push people into categories that are hard to break out (O'Neil, 2016).

3. Methods

In order to better understand the size and nature of gender inferences, an online pilot survey was performed on the accuracy of gender inferences on Twitter. On the basis of this survey and desk research, the implications of inaccurate gender inferences were identified. Given that the historical feminization of gay men (Capsuto, 2000; Russo, 1987), we hypothesized, then, that Twitter was feminizing gay men users more than straight men users.

To validate the hypothesis, we conducted an online pilot survey disseminated using Twitter.[7] For four days, from 22 to 26 May 2020, $N = 109$ Twitter users responded. Tweets have been found to reach their halflife of peak engagement with other users by 18 to 24 min since posting (Bray, 2012; Rey, 2014). Having conducted the pilot survey over four days left a sufficient 'buffer zone' to allow follow-up, less frequent engagement beyond that first 18–24 min. The survey was prepared in Qualtrics and included five specific questions revolving around the topic of whether Twitter algorithms were inferring the gender of users and whether it was correct. In particular we asked what was 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).[8] Although having the information on their website, that Twitter infers users' gender is not apparent to most users, so we gave users instructions on finding their assigned gender on Twitter.[9] We processed anonymous data, and surveyed the adult population.

At the closing of the survey, data was exported, tabulated, and analyzed using Microsoft Excel Spreadsheet Software. The lead author analyzed the survey data. The remaining authors examined the tabulated data and analysis to discuss any discrepancies and ensure the reliability of the results. Understanding, empirically, if Twitter (mis)genders users lays the foundation for exploring the implications of the social media practice of inferring the gender of users. For the discussion section, we refer to privacy and discrimination law, focusing on the impact of online behavioral advertising on inadvertent privacy violations (Wachter, 2020) and the reinforcement of social prejudices.

4. Results

4.1. Data

The data resulting from the online pilot survey are shown in Table 1.

The survey data show that, 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, but not from their sexual orientation. However, the literature is rich in examples of how apparently fair algorithmic designs and categorizations have ulterior and unintended consequences in specific communities (Caliskan et al., 2017; Gomes, Antonialli & Dias-Oliva, 2019; Ito, 2019; Poulsen, Fosch-Villaronga & Søraa, 2020). Our collected data shows that, out of the misgendered Twitter users that we analyzed, only 38% were straight. We notice that only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% of straight women. Individuals that self-reported as bisexuals were also highly misgendered in 25% of the cases for bisexual women and 20% for bisexual men. In the three cases that respondents identified as non-binary, they were wrongly gendered in all cases. These results show that, in our sample, the LGBTQAI+ community and straight women were more often misgendered than straight men. Moreover, women and non-binary are usually more misgendered by Twitter than men.

Although not statistically significant, lesbian and questioning people respondents were less likely to be misgendered, although the numbers are small here, with only 2 lesbian and 3 questioning participants in our sample. It should be noted that one questioning and one lesbian participant answered that they had provided their gender to Twitter, meaning the gender of the one remaining lesbian and two questioning participants had been inferred correctly by Twitter. The findings also show that non-binary participants ($n = 2$) were misgendered, both of which 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% responded 'I do not know' to

[7] *See* https://twitter.com/eduardfosch/status/1263917852484083712?s=20.

[8] The exact wording of the questions was: 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?

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

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

Twitter gender inference accuracy in a $N = 109$ sample 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				0	0%			0%
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%	0%		0%
Other	2		1	3	2			2	100%			67%
Sum	**50**	**55**	**4**	**109**	**9**	**9**	**3**	**21**	**18%**	**16%**	**100%**	**19%**

this question. 42% of those who did not provide Twitter with gender were from the LGBTQAI+ community. Of the 16 participants who had provided their gender, all but one answered 'Yes' about whether or not the gender's appearing on their Twitter profile was correct. That one 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.

Parallel findings resulted from discussions over Twitter, where we shared the online survey. Some respondents openly reported that time ago, Twitter misgendered them, but that now Twitter gendered them correctly - maybe because of their interest in gender equality. Other respondents highlighted that, although having two profiles, the profile that they used the most the gender was wrongly assigned by Twitter. On other occasions, a respondent seemed to point that, although gay, Twitter assigned his gender correctly, while another was *surprised* to be considered 'female' while being a 'male.'

4.2. Limitations

It is clear that, although all the respondents completed the survey fully, this online pilot survey has several limitations. The most critical limitation is obviously 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. The intention of this pilot survey was to confirm whether gender inferences by Twitter contain inaccuracies. However, a much larger and more diverse sample would be needed to provide a comprehensive understanding of misgendering on Twitter and increase statistical power (Cohen, 1988). Yet, on the hypothesis that Twitter was feminizing gay men users more than straight men users, the data shows that 25% of gay men respondents were misgendered compared to 8% of straight men. As such, while the other patterns mentioned concerning other respondent groups may not be systematic, we observe our hypothesis was confirmed.

A second, related limitation may be the limited representativeness of the sample, which seems to over-represent the LGBTQAI+ 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: the LGBTQAI+ community may be overrepresented among Twitter users (Twitter does not provide data on this), in our Twitter networks, or people from the LGBTQAI+ community may have been more inclined to complete the survey, perhaps because the survey topic appealed to them, as it may relate to past experiences or gender stereotyping or misrepresentation, on Twitter or elsewhere. Typically, a larger sample of non-binary and trans people in the dataset could help determine whether these categories face more harm than cisgender people do as a result of misgendering. Also, it would be important that more respondents remember whether they explicitly provided their gender to Twitter. In our pilot survey, 61% did not recall this, rendering it impossible to assess this group's gendering accuracy. This issue could be mitigated by a large sample or asking respondents more explicitly to check this when completing the survey.

A third limitation is obviously that we only focused on Twitter. To better understand the implications of misgendering by online social media platforms, it would be necessary to include other social media providers in a study with a broader scope. This would provide clues on whether Twitter practices are unique or also exist in different social media.

A fourth limitation relates to our choice for desk research, used for identifying the implications of misgendering. The desk research reveals significant implications of inaccurate gendering practices. However, a more in-depth picture could perhaps be obtained by interviewing respondents who have been misgendered, asking them how this made them feel and their impact on their lives. The online survey provides indications for this, but additional interviews may have added value here.

5. Discussion

5.1. Gender bias may propagate in social media

To infer gender, gender classification systems (GCS) make use of gender-stereotype features, such as body movements, physiological and behavioral characteristics, facial features, and language use, to fit newly observed features in inputted data, such as gait, handshape, or sentiment, into either a masculine or feminine category using a trained classifier (Rai & Khanna, 2012). In this sense, gender can be understood as a social construct (Butler, 1990; Zimmerman & West, 1987), which is apparent in our case-study, where

Twitter constructs gender identities of its users whether that corresponds to the gender users' identify with or not. However, there is an intrinsic problem in inferring gender based on stereotypes that assign masculine features to men and feminine qualities to women because not only do these stereotypes create classifier bias in GCS, they also perpetuate gender stereotyping. Since gender identity is primarily subjective and internal, it also clashes with the idea that gender can be recognized automatically, at least with the state of art GCS (Hamidi et al., 2018).

Classifiers trained on real-world datasets are often biased because the data used to train them is biased, containing namely racial and gender stereotypes (Buolamwini & Gebru, 2018; Font & Costa-jussa, 2019; McDuff, Ma, Song & Kapoor, 2019; Torralba & Efros, 2011). For instance, female names are more associated with family than career words, with arts more than mathematics and science (Nosek, Banaji & Greenwald, 2002a, 2002b). Zhao, Wang, Yatskar, Ordonez and Chang (2017) found that the datasets imSitu and MS-COCO are significantly gender-biased and that "models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data" (p. 7). For example, the verb 'cooking' was found to be heavily biased towards females in a classifier trained using the imSitu dataset, amplifying existing gender biases (Zhao et al., 2017). The same gender biases have been shown in natural language processing (Sun et al., 2019; Zhou et al., 2019), another method used to support gender classifiers (Campa, Davis & Gonzalez, 2019). These gender biases in the offline world may propagate to artificial intelligence if not addressed carefully (Caliskan et al., 2017). This is particularly worrisome if we understand that available research suggests that many individuals perceive automatic misgendering as more harmful than human misgendering (Hamidi et al., 2018).

Language is an inherent carrier of socio-linguistic semiotic values, which in English perhaps is less evident than, for instance, in Japanese. Let us consider Japanese Twitter, and how male and female language differ: where the word for "I/myself" in the standard gender-neutral is "watashi" but where female speech has the additional word "atashi" and male speech has words like "boku, ore." Twitter in Japanese would thus potentially be even more prone to gendering as it would know from the gendered language in a linguistic context if the person speaking was more male/female-identifying. Similar linguistic gender-guessing was problematized by Schiebinger (2014) in the case of google translate, making all professors—even herself—into "male" professors when translating English to Spanish. More comparative research in non-English languages is needed to see how linguistic parameters impact twitter gendered assignments.

Although some authors propose the use of *fairness in machine learning* to include non-discrimination mathematical formulations in decision-making (Caliskan et al., 2017), a better approach would be to 1) ask directly Twitter users their gender, and, in case they refuse to disclaim it, 2) do not infer it without the user consent.

5.2. Misgendering has adverse consequences

Misgendering users via automated gender recognition systems also has broader, adverse implications, some of those being that they reinforce gender binarism, undermine autonomy, are a tool for surveillance, and threaten safety (Hamidi et al., 2018). Misgendering is particularly problematic for communities that have been historically discriminated against and for communities which gender is a sensitive part of their identity (McLemore, 2015; Fergus, 2020). Misgendering reinforces the idea that society does not consider a person's gender real, causing rejection, impacting self-esteem and confidence, the felt authenticity, and increasing one's perception of being socially stigmatized (Keyes, 2018).

Several studies and findings in the field of data bias illuminate how data bias and algorithmic gendering affect users from different stands. One such example of data bias is "Statistical discrimination," which refers to making (un)educated guesses about an unobservable candidate characteristic, such as which applicants' perform well as employees. This has been proven quite problematic from the Amazon-hiring algorithm failure, where women candidates were more often devalued than men, as the company traditionally had hired few women (Bogen, 2019). The algorithm concluded that being a woman was an undesirable characteristic for recruitment purposes. Thus, having a CV with the entry of being president of the "women's chess club" was seen as a red-flag, giving the candidate more negative scores, while just generally being a member of a "chess club" was seen as positive. Additionally, "Women on Wikipedia tend to be more linked to men than vice versa. On a lexical level, we find that especially romantic relationships and family-related issues are much more frequently discussed on Wikipedia articles about women than men" (Wagner, Garcia, Jadidi & Strohmaier, 2015, 2016).

When the tools used to extract patterns and profiles from data are not transparent, it may be hard for people to contend any decisions resulting from this, which may impede their freedom and autonomy. On top of that, if sensitive attributes, such as sexual orientation, ethnicity, religion, or trade union membership, are used for decision-making, this may result in discrimination, also from a legal perspective. For instance, in the EU the collecting and processing of personal data is protected under the General Data Protection Regulation (GDPR), which also addressed discrimination issues in datasets (see Recital 71 of the GDPR).[10] However, 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 (Wachter & Mittelstadt, 2019).

This discrimination can be direct or indirect (i.e., by proxy). When dealing with discrimination in (patterns and profiles extracted from) large datasets, proxy discrimination can be hard to detect. Hence, indirect discrimination can occur unintentionally when profile

[10] Recital 71 of the GDPR states 'the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate (...) that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.'

users are unaware of any harm they may be doing. However, it may also be the case that companies use profiles precisely to conceal discrimination, which is referred to as masking (Custers, 2013). Because direct discrimination in data is hard to detect and indirect discrimination is even hard to detect, it can be difficult to enforce equal treatment acts and data protection legislation.

Many forms of discrimination are illegal in most Western jurisdictions (for instance, when hiring people based on gender or ethnicity), or at least controversial (for instance, when not hiring people because they have a criminal record). Legislation that bans discrimination on the grounds of specific characteristics concerns, on the one hand, lists of characteristics that may not serve as a basis for making decisions (including gender, ethnicity, political preferences, trade union membership, or sexual orientation); on the other hand, certain types of decisions are forbidden (such as hiring and firing employees or offering products and services). Not every decision based on the sensitive characteristics mentioned is forbidden, however. For example, it is a matter of personal choice who your friends are. Nonetheless, 'weaker' forms of discrimination may occur in the formation of friendships, in the form of stigmatization of specific population groups. On a larger scale, this could lead to social polarization and segregation. For now, using wrong pronouns or misgendering is not sufficiently grave to be considered harassment under certain specific legal provisions, although several parts of the population may advocate for remedies that make justice to these acts (Ashley, 2017).

5.3. *Gender may be more sensitive than we thought*

People and devices currently generate large amounts of data via social media, websites, trackers, and sensors in devices often connected to the Internet of things (IoT). Some of the data is generated explicitly by people, for instance, when they post messages on social media and create websites. However, some data is generated beyond the knowledge and awareness of people, for instance, when devices collect data in the background and when devices communicate with each other. In order to make sense of the vast amounts of available data, many organizations, companies, and governments alike try to find patterns and profiles in such large datasets, usually via tools like machine learning and data mining (Custers & Bachlechner 2018). Profiling techniques like regression, classification, or clustering, mainly ascribe attributes to people (Calders & Custers 2013). These techniques infer specific people's attributes from different inputs of data, coming either from the same person (i.e., estimating recidivism based on someone's criminal record) or others (i.e., others who ordered this book, also like these books).

A critical feature of automated recognition systems is that companies infer information from data not directly or indirectly provided by data subjects (Custers, 2018). Twitter makes inferences about users' accounts, including interests, age, and gender of users to provide features such as account suggestions (e.g., suggested contacts, promoted accounts for the user to follow), advertising, recommendations, and timeline ranking.[11] Twitter makes use of users' content, activity, relationships, and interactions to genderize content production patterns (Robinson et al., 2015), infer gender, and make these suggestions.[12] 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, it is the trade-off that users have to accept if they want to have a personalized Twitter account.

As a legal basis, Twitter states that it makes "inferences about your account - such as interests, age, and gender" for "legitimate purposes."[13] The appeal to legitimate interests is controversial. The GDPR lists a limited number of legal grounds for data processing, including consent, performance of a contract or legitimate interests. Although the legitimate interests ground should not be considered as a 'last resort' when all other grounds for legitimate data processing fail, it should not be automatically chosen or its use unduly extended on the basis of a perception that it is less constraining than other grounds (A29WP, 2014). Legitimate purposes is only a solid legal basis for data processing if there is a necessity. It is questionable whether these inferences are necessary for Twitter.

According to the UK's Information Commissioner's Office (ICO, 2020), 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. Moreover, they enunciate that if controllers choose this legal ground, they 'are taking on extra responsibility for considering and protecting people's rights and interests' (ICO, 2020). ICO also points out three elements to the basis of the legitimate interest: identify a legitimate interest; show that the processing is necessary to achieve it; and balance it against the individual's interests, rights, and freedoms.

However, our survey results highlight many misgendered users and question whether Twitter did balance their interests against individuals' interests. First, of the 109 participants, only 15% provided Twitter with their gender, while Twitter inferred their gender as part of the personalized. Second, the results seem to indicate that some LGBTQAI+ community members and straight women are 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 from that inference unless she unticks the personalization box. Making the user choose between these two is as if in times of COVID-19, developers made users choose between health or privacy (Harari, 2020). It results in something of 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 (Custers, 2013).

People have a right to access their data (Art. 15 of the GDPR) and receive meaningful information about the logic involved in the

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

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

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

data analytics (Art. 22 of the GDPR). In the case of Twitter, users can check their inferred gender under 'your Twitter data.' However, this process is often opaque to data subjects, who largely ignore that Twitter inferred such an attribute. If they have access, they have no right to access the algorithms and data of other data subjects used in the analyses, as companies may see those algorithms and profiles as trade secrets of their vital interest (Custers, 2018). In such a case, they may not be able to check whether inferred data is correct. The logic involved in machine learning algorithms may hamper the enforcement of other rights, such as rectification or the right to be forgotten (Fosch-Villaronga, Kieseberg, & Li, 2018).

5.4. *Algorithms should account for diversity and inclusion*

Misgendering users in the background is not good practice, and can lead to privacy and discrimination issues, beyond echoing deeply rooted stereotypes. In this sense, platforms may be excluding and misrepresenting a large number of potential users if they are not respectful and inclusive towards their gender identity or sexual orientation. As Fergus (2020) announced, 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). These findings may result from the fact that twitter GCS do not account for diversity and operate with a male/female binary categorization that does represent non-binary people's gender expression and does no justice to the freedom of identity that everyone should have.

From the results of our short online survey it is apparent 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. By being misgendered, one seems to be essentially told that something is not "correct" with the way one acts, particularly online. In this sense, the lack of diversity can lead to toxic cultures and algorithmic bias (Lambrecht & Tucker, 2018). From all this, it is clear that digital identity and participatory culture plays a massive role in the sense of self in the modern world and that there should be more efforts towards realizing diversity and inclusion in the online world (Jenkins, 2015) to not perpetuate the normative view that certain collectives such as trans or non-binary do not exist (Keyes, 2018).

Machine learning and data mining tools can be developed in such a way that they do not yield discriminating patterns, such as gender-based patterns or profiles (Kamiran, Calders & Pechenizkiy, 2013). This approach is referred to as discrimination-aware data mining. The underlying idea is not to restrict 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 for indirect dsicrimination (discrimination by proxy). Focusing on the design of the algorithms can prevent this, when using gender in the development of data-driven decision models (Zliobaite & Custers 2016).

Better accounting for diversity and inclusion earlier on in the gender-targeted advertising and content suggestion ecosystems could reduce bias in other systems using GCS in these areas. For example, take a recommendation system on social media platforms that use characteristics, including gender, to group users, and recommend social network groups to users (Baatarjav, Phithakkitnukoon & Dantu, 2008). Using potentially inaccurate inferences made with a bias GCS, a group recommendation system might recommend a misgendered lesbian woman join a male-targeted social network group. Having a GCS that accounts for diversity and inclusion would help reduce bias in systems in which gender inferences flow, including search and recommendation systems, which similarly need to be *fairness-aware* (Geyik, Ambler & Kenthapadi, 2019).

Another way to account for diversity in this area may be to use algorithms to detect and quantify the extent to which search engines, social media, or other platforms respond to stereotypically gendered queries containing stereotypical language. In this respect, some authors have proposed what they call a *Gender Stereotype Reinforcement measure*, which quantifies search engines' trend to promote gender stereotypes, leveraging gender-related information encoded in word embeddings to counterbalance gender bias (Fabris, Purpura, Silvello & Susto, 2020). These efforts align somewhat with the aims of content moderator tools, i.e., detect potential offensive and unwanted images, filter possible profanity and undesirable text, or moderate adult and racy content in videos.[14] However, these tools do not come without drawbacks. For instance, text-based tools similar to the proposed by Fabris et al. (2020) may miss important contextual nuances difficult to be detected objectively in written language (Gomes et al., 2019).

6. Conclusions and future work

In this article, we have addressed the research question: *What are the implications of inaccurate gender inferences by Twitter?* Since the implications of inaccurate gender inferences can only be identified if it is clear how the gender inferences work and how (in)accurate they actually are, an online pilot survey to ascertain the accuracy of Twitter's gender inferences was conducted. The data collected via this survey shows that, out of $N = 109$ respondents, Twitter inferred users' gender in the majority of cases. Concerning the accuracy, 81% of the cases were accurate, while 19% were misgendered. Only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% of straight women. Non-binary users were misgendered in all the cases.

Based on these results, we identified four major implications of inaccurate gender inferences. First, using both accurate and inaccurate gender inferences may propagate gender bias in social media. Second, misgendering has adverse consequences, both clear and less obvious ones. The adverse consequences can apply to individuals (including their social position and personal feelings) and groups (including discrimination and stigmatization). Third, it is becoming increasingly clear that gender may be more sensitive than

[14] *See* for instance https://azure.microsoft.com/en-us/services/cognitive-services/content-moderator/.

we thought, in the sense that inaccuracies can have worse than anticipated consequences. Fourth, considering the gendering practices, it increasingly becomes clear that dealing with this requires that algorithms be modified to account for diversity and inclusion.

Social media platforms like Twitter have economic incentives to know users' genders for commercialization and targeted advertisements. In this way, both attributes that a data subject may not want to disclose and attributes that a data subject does not know can be predicted via data analytics can be ascribed to that person. Our investigation shows, however, that the inferred gender may be inaccurate and morally questionable. To exemplify why misgendering in social media is an issue, consider the miscalculation of other personally identifiable traits in everyday interactions. Misreading users' sexual orientation, race, age, or ableness would probably meet with considerable resistance. Although data accuracy for behavioral advertising is susceptible to errors, not accounting for diversity and inclusion may render online behavioral advertising inefficient and lead to serious privacy, discrimination, autonomy, and self-identity issues (Keyes, 2018).

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. Automated gender recognition systems should therefore account for diversity and inclusion, using a more accurate understanding of gender identity to fully represent contemporary society – if indeed needing to know a users' gender is important in the first place. Otherwise, inferential analytics may reinforce existing biases about gender stereotyping and have adverse consequences on many parts of society. 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 fair. Inclusive engagement that reflects on the users as not homogeneous can have a positive impact on technology.

With this said, it is also important to notice that, paradoxically, technology increasingly has a habit of being both the source of and the solution to societal problems, and algorithms are no exceptions to that (Bauman, 2013; Fosch-Villaronga, 2019). There are countless examples of how technology has been proposed to solve inadequate engineering practice, government policy failures, or modern consumerism outcomes, showing how technological fixes have cultural, ethical, and political implications (Bauman, 2013; Johnston, 2018). Automated gender recognition systems may offer a good solution to recognize gender automatically for many applications, but these systems misgender users. Some researchers work on tools to counter gender bias (Fabris et al., 2020), but despite excellent intentions these propositions also miss the most fundamental aspect of gender: that gender is subjective, and that gender cannot be objectively recognized. In Johnston's (2018) words, "modern problems cannot be reduced to mere engineering solutions over the long term; human goals are diverse and constantly changing."

Understanding the impact that algorithms have on different communities is challenging, as they may appear much later, usually after being widely used (Hao, 2019b), but undoubtedly necessary to make a fairer society. It is often the case that those communities mostly remain "invisible, silent, powerless, and unable to understand how these technologies may affect them" (Poulsen et al., 2020). Our study is one step closer to giving voice to these communities by showing that social media misgendering may affect communities such as the LGBTQAI+ community and straight women, that have historically suffered from discrimination (Søraa et al., 2020).

Before starting this research, a larger scale study did not seem justified since we only had limited indications (some literature and a statement on Twitter). However, given the results we found, as future work, we intend to conduct a more extensive and refined survey to investigate this issue together with the user's impressions. With a more diverse sample, a larger survey ought to add further to our understanding of the problem of misgendering on Twitter. For example, since non-binary and trans people face more harm as a result of misgendering than cisgender people do, a larger sample of non-binary and trans people in the dataset was expected. Looking forward, it is also important to consider other *classifications* that technological and bureaucratic procedures generally materialize due to preexisting and prevalent biases and prejudices, inequalities, and power structures, including social class, geographic space (Katzenbach & Ulbricht, 2019).

By identifying how inferential analytics may reinforce gender stereotyping and affect marginalized communities, we hope to continuously contribute to promoting the need for privacy, diversity, and inclusion online and advocate for the freedom of identity that everyone should have online and offline.

CRediT authorship contribution statement

E. Fosch-Villaronga: Conceptualization, Investigation, Methodology, Resources, Writing - original draft, Writing - review & editing. **A. Poulsen:** Investigation, Methodology, Resources, Writing - review & editing. **R.A. Søraa:** Investigation, Methodology, Resources, Writing - review & editing. **B.H.M. Custers:** Investigation, Resources, Supervision, Writing - review & editing.

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.

Annex

The data used for this study is available following this link: https://docs.google.com/spreadsheets/d/1qmHX0mktQQZ50VQFFlj6K9_MrL7W_zREEEXk6_RYIi4/edit?usp=sharing.

References

Article 29 Working Party, A29WP (2014). Opinion 06/2014 on the notion of legitimate interest of the data controller under Article 7 of Directive 95/46/EC. 844/14/EN, WP 217, adopted 9 April 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

Ashley, F. (2017). 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 (accessed 30 May 2020).

Baatarjav, E. A., Phithakkitnukoon, S., & Dantu, R. (2008). Group recommendation system for Facebook. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *5333. On the Move to Meaningful Internet Systems: OTM 2008 Workshops. OTM 2008. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-88875-8_41.

Bauman, Z. (2013). *Liquid love: On the frailty of human bonds*. Cambridge, UK: John Wiley & Sons.

Beauvoir, S. D. (1949). *Le deuxième sexe*. Paris: Éditions Gallimard.

Bogen, M. (2019). *All the ways hiring algorithms can introduce bias*. Harvard Business Review. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias (accessed 30 May 2020).

Bray, F. (2007). Gender and technology. *Annual Review of Anthropology, 36*, 37–53.

Bray, P. (2012). When Is My Tweet's Prime of Life? (A brief statistical interlude.). *Moz*. Available at https://moz.com/blog/when-is-my-tweets-prime-of-life (accessed 2 September 2020).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91.

Burdge, B. J. (2007). Bending gender, ending gender: Theoretical foundations for social work practice with the transgender community. *Social Work, 52*(3), 243–250.

Butler, J. (1990). Gender trouble, feminist theory, and psychoanalytic discourse. *Feminism/postmodernism*, 327.

Calders, T., & Custers, B. H. M. (2013). What is data mining and how does it work? In B. H. M. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and privacy in the information society*. Heidelberg: Springer.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.), 356*(6334), 183–186.

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

Capsuto, S. (2000). *Alternate channels: The uncensored story of gay and lesbian images on radio and television*. New York City: Ballantine Books.

Cohen, Jacob (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). Hillsdale, New Jersey: L. Erlbaum Associates.

Corney, M., De Vel, O., Anderson, A., & Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. In *, 2002. Proceedings of the 18th IEEE Annual Computer Security Applications Conference* (pp. 282–289).

Custers, B. H. M. (2013). Data Dilemmas in the Information Society. In B. H. M. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and privacy in the information society*. Heidelberg: Springer.

Custers, B. H. M. (2018). Profiling as inferred data: Amplifier effects and positive feedback loops. (red.) In E. Bayamlioglu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *Being profiled: Cogitas ergo sum: 10 years of "profiling the European citizen"* (pp. 112–116). Amsterdam: Amsterdam University Press.

Custers, B. H. M., & Bachlechner, D. (2018). Advancing the EU data economy: Conditions for realizing the full potential of data reuse. *Information Polity, 22*(4), 291–309.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy*. Boston: Harvard Business School Press.

Fabris, A., Purpura, A., Silvello, G., & Susto, G. A. (2020). Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management, 57*(6), Article 102377.

Fergus, J. (2020). 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 (accessed 30 May 2020).

Filho, Lopes, Ahirton, José, & Pasti, Rodrigo, & De Castro, Leandro (2016). Gender classification of Twitter data based on textual meta-attributes extraction. 10.1007/978-3-319-31232-3_97.

Fink, C., Kopecky, J., & Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. In *Sixth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence (AAAI)* (pp. 459–462). Available at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4644/5031 (accessed 28 May 2020).

Font, J.E., & Costa-jussa, M.R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. Available at https://arxiv.org/pdf/1901.03116.pdf (last accessed 12 February 2021).

Fosch-Villaronga, E. (2019). *Robots, healthcare, and the law: Regulating automation in personal care*. New York City: Routledge.

Fosch-Villaronga, E., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: Artificial intelligence and the rightto be forgotten. *Computer Law & Security Review, 34*(2), 304–313.

Garibo-Orts, O. (2018). A Big Data approach to gender classification in Twitter. In *Proceedings of PAN at CLEF 2018. Avaialble at*. http://ceur-ws.org/Vol-2125/paper_204.pdf (accessed 25 May 2020).

Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). *Racial and gender discrimination in transportation network companies (No. w22776)*. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w22776/w22776.pdf (last accessed 12 February 2021).

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search.. Paper presented at the. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3292500.3330691.

Glick, P., Gangl, C., Gibb, S., Klumpner, S., & Weinberg, E. (2007). Defensive reactions to masculinity threat: More negative affect toward effeminate (but not masculine) gay men. *Sex roles, 57*(1–2), 55–59.

Gomes, A., Antonialli, D., & Dias-Oliva, T. (2019). Drag queens and Artificial Intelligence: Should computers decide what is 'toxic' on the internet? *Internet Lab Blog*. https://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/ (accessed 4 June 2020).

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

Haas, L., & Hwang, C. P. (2007). Gender and organizational culture: Correlates of companies' responsiveness to fathers in Sweden. *Gender & Society, 21*(1), 52–79.

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (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* (pp. 1–13).

Hänold, S. (2018). Profiling and Automated Decision-Making: Legal Implications and Shortcomings. In M. Corrales, M. Fenwick, & N. Forgó (Eds.), *Robotics, ai and the future of law. perspectives in law, business and innovation* (pp. 123–153). Singapore: Springer.

Hao, K. (2019a). 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/ (accessed 30 May 2020).

Hao, K. (2019b). This is how AI bias really happens-and why it's so hard to fix. *MIT Technology Review*. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/ (accessed 30 May 2020).

Harari, Y. N. (2020). The world after coronavirus. *Financial Times*, 19 March. Available at: https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75 (accessed 30 May 2020).

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). 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), 1–19.

Hildebrandt, M., & Gutwirth, S. (2008). *Profiling the European citizen*. Heidelberg: Springer.

Hooper, C. (2001). *Manly states: Masculinities, international relations, and gender politics*. New York City: Columbia University Press.
Howansky, K., Wilton, L. S., Young, D. M., Abrams, S., & Clapham, R. (2019). Trans) gender stereotypes and the self: Content and consequences of gender identity stereotypes. *Self and Identity*, 1–18.
ICO (2020). 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/ (accessed 30 May 2020).
Ito, J. (2019). Supposedly 'fair' algorithms can perpetuate discrimination. *MIT Media Lab*. Retrieved from https://www.media.mit.edu/articles/supposedly-fair-algorithms-can-perpetuate-discrimination/ (accessed 30 May 2020).
Jenkins, H., & Ito, M. (2015). *Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics*. Hoboken, New Jersey: John Wiley & Sons.
Jernigan, C., & Mistree, B. F. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday, 14*(10).
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
Johnston, S. F. (2018). *The technological fix as social cure-all: Origins and implications, 37* pp. 47–54). IEEE Technology and Society Magazine.
Kachel, S., Steffens, M. C., & Niedlich, C. (2016). Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in Psychology, 7*(956), 1–19.
Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In B. H. M. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *(2013) discrimination and privacy in the information society* (pp. 223–239). Heidelberg: Springer.
Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review, 8*(4). https://doi.org/10.14763/2019.4.1424.
Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. In *, 2. Proceedings of the ACM on Human-Computer Interaction* (pp. 1–22).
Khan, S. A., Ahmad, Munir, Nazir, Muhammad, & Riaz, N. (2014). A comparative analysis of gender classification techniques. *Middle - East Journal of Scientific Research, 20*, 1–13. https://doi.org/10.5829/idosi.mejsr.2014.20.01.11434.
Klein, E. (2013). *Gender politics: From consciousness to mass politics*. New York City: Harvard University Press.
Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences, 110*(15), 5802–5805.
Kumar, Deepak, Gupta, Rajat, Sharma, Ashirwad, & Kumar Saroj, Sushil (2019). Gender Classification using Skin Patterns. In *March 12,. Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019. Available at SSRN: https://ssrn.com/abstract=3351003.
Lambrecht, A., & Tucker, C.E. (.2018). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260 (accessed 30 May 2020).
Li, B., Lian, X.-. C., & Lu, B.-. L. (2012). Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing, 76*(1), 18–27. https://doi.org/10.1016/j.neucom.2011.01.028.
Lin, Feng, & Wu, Yingxiao, & Zhuang, Yan, & Long, Xi, & Xu, Wenyao. (2015). Human gender classification: A review. http://cse.ucdenver.edu/~linfen/papers/2016_IJBM_gender.pdf.
Mathivanan, P., & Poornima, K. (2018). Biometric authentication for gender classification techniques: A review. *Journal of The Institution of Engineers (India): Series B, 99*, 79–85. https://doi.org/10.1007/s40031-017-0299-z.
Matz, S. C., Menges, J. I., Stillwell, D. J., & Schwartz, H. A. (2019). Predicting individual-level income from Facebook profiles. *PloS one, 14*(3).
McDuff, D., Ma, S., Song, Y., & Kapoor, A. (2019). Characterizing bias in classifiers using generative models. arXiv preprint 1906.11891 https://arxiv.org/abs/1906.11891.
McLemore, K. A. (2015). Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity, 14*(1), 51–74.
Nature. (2020). Accounting for sex and gender makes for better science. *Editorial, Nature*. Retrieved from https://www.nature.com/articles/d41586-020-03459-y (last accessed 6 January 2020).
Nieuwenhuis, M., & Wilkens, J. (2018). Twitter text and image gender classification with a logistic regression n-gram model. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*.
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York City: NYU Press.
Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115.
Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology, 83*(1), 44–59.
Office of the High Commissioner Human Rights. (2020). *Gender stereotypes*. New York City: United Nations. Available at https://www.ohchr.org/EN/Issues/Women/WRGS/Pages/GenderStereotypes.aspx (last accessed 25 May 2020).
Ogasawara, Y. (1998). *Office ladies and salaried men: Power, gender, and work in Japanese companies*. Oakland: Univ of California Press.
O'Neil, C. (2016). *Weapons of math destruction; how big data increases inequality and threatens democracy*. New York: Crown.
Park, Sunghee, & Woo, Jiyoung (2019). Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum. *Applied Sciences., 9*, 1249. https://doi.org/10.3390/app9061249.
Pinsof, D., & Haselton, M. G. (2017). The effect of the promiscuity stereotype on opposition to gay rights. *PloS one, 12*(7), Article e0178534.
Poulsen, A., Fosch-Villaronga, E., & Søraa, R. A. (2020). Queering machines. *Nature Machine Intelligence, 2*(3), 152-152.
Rai, P., & Khanna, P. (2012). Gender classification techniques: A review. In D. Wyld, J. Zizka, & D. Nagamalai (Eds.), *166. Advances in Computer Science, Engineering & Applications. Advances in Intelligent and Soft Computing*. Berlin, Heidelberg: Springer.
Randall, V., & Waylen, G. (Eds.). (2012). *Gender, politics and the state*. New York City: Routledge.
Rey, B. (2014). Your tweet half-life is 1 billion times shorter than Carbon-14's. *Wiselytics*. Available at http://www.wiselytics.com/blog/tweet-isbillion-time-shorter-than-carbon14/ (accessed 2 September 2020).
Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., et al. (2015). Digital inequalities and why they matter. Information. *Communication & Society, 18*(5), 569–582. https://doi.org/10.1080/1369118X.2015.1012532.
Roosendaal, A. (2010). Facebook Tracks and Traces Everyone: Like This! Tilburg Law School Legal Studies Research Paper Series No. 03/2011. Available at SSRN: https://ssrn.com/abstract=1717563.
Rosa, P., & Dawson, A. (2006). Gender and the commercialization of university science: Academic founders of spinout companies. *Entrepreneurship and Regional Development, 18*(4), 341–366.
Russo, V. (1987). *The celluloid closet: Homosexuality in the movies*. Amsterdam: Harper Collins.
Schiebinger, L. (2014). Scientific research must take gender into account. *Nature, 507*(7490), 9-9.
Sink, A., Mastro, D., & Dragojevic, M. (2018). Competent or warm? A stereotype content model approach to understanding perceptions of masculine and effeminate gay television characters. *Journalism & Mass Communication Quarterly, 95*(3), 588–606.
Søraa, R. A. (2017). Mechanical genders: how do humans gender robots? *Gender, Technology and Development, 21*(1-2), 99–115.
Søraa, R. A., Anfinsen, M., Foulds, C., Korsnes, M., Lagesen, V., Robison, R., et al. (2020). Diversifying diversity: Inclusive engagement, intersectionality, and gender identity in a European Social Sciences and Humanities Energy research project. *Energy Research & Social Science, 62*, 1–11.
Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., et al. (2019). *Mitigating gender bias in natural language processing: Literature review* (pp. 1630–1640). https://doi.org/10.18653/v1/P19-1159. https://arxiv.org/pdf/1906.08976.pdf (accessed 12 February 2021).
Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J., & Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature, 575*(7781), 137–146.
Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2019). Algorithmic inference, political interest, and exposure to news and politics on Facebook. Information. *Communication & Society*, 1–18. https://doi.org/10.1080/1369118X.2019.1642934.
Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1521–1528). IEEE.

United Nations Human Rights Office of the High Commissioner, OHCHR. (2012). *The corporate responsibility to respect human rights. an interpretative guide.* New York City: United Nations. Available at https://www.ohchr.org/Documents/publications/hr.puB.12.2_en.pdf (accessed 1 September 2020).

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the ACM Symposium On Usable Privacy and Security (SOUPS), July 11-13, 2012* (pp. 1–15).

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

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

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*. https://arxiv.org/pdf/1501.06307.pdf.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in Wikipedia. *EPJ Data Science, 5*(1), 5.

Wilchek-Aviad, Y., Tuval, C., & Zohar, N. (2020). Gender stereotyping and body image of transgender women. *Current Psychology*, 1–10.

Willson, M. (2017). Algorithms (and the) everyday. Information. *Communication & Society, 20*(1), 137–150. https://doi.org/10.1080/1369118X.2016.1200645.

Yan, Xiang, & Yan, Ling (2006). Gender Classification of weblog authors. In *Proceedings of the Conference: Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006* (pp. 228–230).

Zarsky, T. (2003). Mine your own business! Making the case for the implications of the data mining of personal information in the forum of public opinion. *Yale Journal of Law and Technology, 5*, 57.

Zhao, J., Wang, T., Yatskar, M, Ordonez, V., & Chang, K. W. (2017). *Men also like shopping: Reducing gender bias amplification using corpus-level constraints* (pp. 2979–2989). https://arxiv.org/pdf/1707.09457.pdf (last accessed 12 February 2021).

Zhou, Pei, Shi, Weijia, Zhao, Jieyu, Huang, Kuan-Hao, Chen, Muhao, Cotterell, Ryan, et al. (2019). *Examining gender bias in languages with grammatical gender* (pp. 5279–5287). https://doi.org/10.18653/v1/D19-1531.

Zimmerman, D. H., & West, C. (1987). Doing gender. *Gender and Society, 1*(2), 125–151.

Zliobaite, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, (24), 183–201.

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