Universiteit
Leiden
The Netherlands

User perceptions of misgendering algorithms

Fosch Villaronga, E.; Mut Piña, A.; Verhoef, T.; Poulsen, A.; Søraa, R.A.; Drukarch, H.G.; ... ; Custers, B.H.M.

Citation
Fosch Villaronga, E., Mut Piña, A., Verhoef, T., Poulsen, A., Søraa, R. A., Drukarch, H. G., ... Custers, B. H. M. (2025). User perceptions of misgendering algorithms. *Big Data &Amp; Society*, *12*(4). doi:10.1177/20539517251398719

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/4288125

Note: To cite this publication please use the final published version (if applicable).

Check for updates

Original research article

BIG DATA & SOCIETY

Big Data & Society
October–December: 1–15
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517251398719
journals.sagepub.com/home/bds
Sage

User perceptions of misgendering algorithms

Eduard Fosch-Villaronga[1], Antoni Mut-Piña[1], Tessa Verhoef[2], Adam Poulsen[3], Roger A. Søraa[4], Hadassah Drukarch[1], Anne Noel[1], Scarlet Tunney[1] and Bart Custers[1]

Abstract

Gender classification systems (GCSs) on social media platforms, such as X (formerly Twitter), infer users' gender for targeted advertising and personalization. However, these systems rely on binary classifications that fail to capture gender diversity, often leading to misclassification (i.e. misgendering). This study is a comprehensive analysis of algorithmic misgendering and its impact on user perceptions through a large-scale online global survey (N = 1523). In the first stage, we assess the prevalence of misgendering on X by analyzing the accuracy of inferred gender classifications. Our findings reveal that women and LGBTQ+ users are disproportionately misclassified, which highlights structural biases in gender inference systems. In the second stage, given that the emotional and social consequences of algorithmic gender inference remain underexplored, we examine how users perceive and respond to misgendering. Using ordinal logistic regression models, we find that individuals who experience misgendering report significantly more aversion to X's gender policies. Furthermore, increased platform engagement is linked to stronger opinions on gender inference, reducing neutrality toward these systems. Beyond these findings, our study also reveals how opaque gender classification is, as many users struggle to locate, understand, or challenge their inferred gender within X's interface. This lack of transparency raises concerns about agency, algorithmic literacy, data protection, and fairness. Based on our work, we suggest regulatory measures to ensure greater transparency and user control over gender classification, contributing to ongoing debates on algorithmic discrimination and inclusive AI governance on current and future social media platforms.

Keywords

Gender, social media, gender classification system, algorithmic bias, discrimination, transparency

Introduction

Gender classification systems (GCSs) on social media platforms, such as X (formerly Twitter), infer users' gender for targeted advertising and personalization (Argamon et al., 2007; Burger et al., 2011; Ur et al., 2012). These systems apply computational techniques to classify gender based on text, images, or behavioral data. Despite advancements in this technology's accuracy, however, GCSs typically operate within a binary framework, failing to account for the complexity and fluidity of gender identities (Keyes, 2018; Schroeder, 2021). As a result, GCSs can misclassify users, reflecting and potentially exacerbating existing biases (Fosch-Villaronga et al., 2021). Although efforts like those by Fabris et al. (2020) aim to mitigate these gender biases, they often overlook that the inherent subjectivity of gender identity cannot be captured through objective means. Echoing Johnston's work (2018), this modern challenge necessitates solutions beyond mere technological fixes (Poulsen et al., 2020; Vásárhelyi and Brooke, 2025).

At the technical level, automated gender inference consistently produces unequal error rates across different genders and LGBTQ+ identities (Ruberg and Ruelos, 2020). Large-scale audits of Twitter's text-based classifiers have revealed higher misclassification rates for women attributed to the overrepresentation of male-coded language in training corpora

[1]eLaw, Center for Law and Digital Technologies, Leiden University, Leiden, Netherlands
[2]Creative Intelligence Lab, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, Netherlands
[3]Brain and Mind Centre, University of Sydney, Sydney, NSW, Australia
[4]Center for Technology and Society, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Corresponding author:
Roger Andre Søraa, Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), 6430A Bygg 6 Dragvoll, Trondheim, Norway.
Email: roger.soraa@ntnu.no

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

(Bivens and Haimson, 2016; Keyes, 2018). Non-heterosexual and non-binary users are disproportionately misgendered because existing models struggle with queer self-presentation online (Fosch-Villaronga et al., 2021; Hamidi et al., 2018). Likewise, gender expressions that depart from cis-normative binaries—such as androgynous avatars or mixed-gender pronouns—systematically increase error rates (Schroeder, 2021). This reflects a broader tension in machine learning: systems trained on binary labels often underperform when encountering non-binary identities, resulting in compounded harm and unfair outcomes (Fosch-Villaronga and Malgieri, 2024; Wiegand et al., 2024). Given that X relies on comparable textual, visual, and network-based features, we expect these patterns to replicate. We, therefore, hypothesize that men experience statistically fewer misclassifications than women (H1); straight respondents are less likely to be misclassified than non-straight respondents (H2) and that gender non-conforming users are more likely to be misclassified (H3).

Misgendering is not merely a statistical error—it is experienced as symbolic violence that undermines users' dignity, autonomy, and inclusion (Costanza-Chock, 2020; Fergus, 2020, McLemore, 2015). When users—particularly women and LGBTQ+ individuals—are misclassified or observe biased outcomes, they are more likely to perceive the system as exclusionary and unjust (Criado-Perez, 2019; Hamidi et al., 2018; Pino and Edmonds, 2024), which can erode trust in the platform's ethical orientation. These dynamics motivate our hypotheses H4–H7, which examine how identity, engagement, and experiences of misgendering shape user sentiment and perceived trustworthiness of platform governance.

These hypotheses bridge technical audit studies with user-centered research, offering a novel perspective on how algorithmic bias is not only measured in terms of error rates but also felt in terms of exclusion, injustice, and loss of trust. To do so, we conducted a large-scale survey (N = 1523) to assess the extent of algorithmic misgendering. We reported the results in Fosch-Villaronga et al. (2026) and here we present the analysis of its emotional and perceptual impacts on users.

This article is structured as follows: Section 2 reviews the literature on GCSs and user rights. Section 3 presents our methodology, including survey design and statistical modeling. Section 4 reports the results. Section 5 discusses implications for platform governance and user autonomy. Section 6 concludes.

Background information

Gender classification systems and their binarism

A gender classification system (GCS) is a computational model designed to categorize individuals into male or female categories based on various data sources. These systems are widely used across digital platforms, including social media, advertising, and security applications. Usually, they take as a parameter the biological sex of the user (the biological sexual traits apparent or deduced at birth) (Karkazis, 2019) rather than their gender identity (the perceptions that users have of themselves) (Fosch-Villaronga et al., 2021, 2026). Currently, X mentions on their website: "Gender. If you haven't already specified a gender, this is the one associated with your account based on your profile and activity. This information won't be displayed publicly."

The methods used to infer gender vary (Bivens and Haimson, 2016), but common techniques include text-based analysis, where algorithms examine word choices and phrasing to categorize users (Liu and Ruths, 2013; Mueller and Stumme, 2016), as well as image-based analysis, which applies facial recognition to profile pictures (Al Zamal et al., 2012). Additionally, social network structures are leveraged, as the gender of a user's connections is often used as a predictive factor based on the assumption that individuals tend to interact within gendered clusters (Chen et al., 2015). Many GCSs integrate multiple inference methods to refine classification accuracy (Barlas et al., 2021), combining textual, visual, and behavioral data to reinforce predictions (Sakaki et al., 2014). By examining these patterns, gender inference algorithms determine the probable gender of a user. Predictions can be based on interactions, such as which tweets a user regularly likes, what search results they engage with, or the time spent on specific threads.

The problem is that this approach stereotypically assumes that certain words, topics, or social connections correspond inherently to male or female identities, a notion deeply rooted in cultural stereotypes and historical biases (Chen et al., 2015; Sakaki et al., 2014). For example, if the dataset predominantly links certain professions with a specific gender—such as "mechanic" with men and "secretary" with women—the algorithm is likely to reinforce those associations. Or, it can take men as the human standard, making women in the data less visible (Criado-Perez, 2019). Datasets may rely on proxy features like names, clothing styles, colors, or activities, as these are quantifiable and frequently correlate with gender in stereotypical ways. Although it is difficult to assess these coded assumptions, as the actual code source of X is not publicly available, it is essential to recognize and investigate what indicators give rise to these gendered inferences.

In this sense, a core limitation of GCSs is their rigid binary framework, which fails to account for non-binary, gender-fluid, and other diverse identities (Alvarado Garcia et al., 2025). Instead of recognizing gender as fluid and socially constructed, these systems operate under the assumption that gender is fixed, universally applicable, and easily categorized (Vivienne et al., 2023). This oversimplification not only excludes individuals who do not conform to traditional gender binaries but also reinforces outdated gender norms through algorithmic decision-making. This risks disproportionately misgendering those with diverse gender and sexual identities, such as those within the broader LGBTQ+[1] community that incorporates a plurality of identities and experiences (Keyes, 2018). Another pitfall is the risk of relying on "other"

categorizations for LGBTQ+ individuals (Lorenz, 2021). Moreover, as GCSs are typically optimized for statistical accuracy rather than fairness, their outputs prioritize alignment with dominant patterns in training data rather than reflecting the diversity of human identities.

The myth of neutrality of gender classification systems

Social media platforms are not passive, neutral infrastructures but active sociotechnical systems that shape and are shaped by user interactions (Verbeek, 2012). Although their digital infrastructure comprises algorithms, AI models, servers, and software and hardware, their operation relies precisely on the social aspect of the technology—people's desire to connect and share stories, ideas, and opinions. However, several studies warn of the risks of echo chambers and the politicization of content due to the reinforcement effects of likes and interactions on these platforms (e.g. Cinelli et al., 2021; Garimella et al., 2018). This contradicts the perception that algorithms function in a neutral, objective manner—instead, they actively reinforce and shape identities, opinions, and experiences.

As a technology, social media has significant potential to influence its users. When viewed as a sociotechnical system (Bijker, 1994), social media have also faced criticism for their effects on users' well-being (Ellison et al., 2022). At the same time, they have opened up new ways of (re)thinking of social media as a sociotechnical system (Ruppert, 2018). Within the sociotechnical system that constitutes social media, identity—including gender identity—is a fundamental part of how the technology operates. Several studies, particularly within feminist technoscience, highlight how technology and gender are co-dependent variables that mutually shape one another (e.g. Bray, 2007; Cockburn and Ormrod, 1993; Rentetzi, 2023; Søraa and Bruijning, 2024). This dynamic is evident in how users' likes influence the content they are subsequently shown. The algorithmic logic reinforces identity-shaping techno-agency; for example, if a user likes car-related content, the algorithm deems it optimal to display more car-related content. Similarly, political likes on either end of the spectrum lead to increased exposure to similar perspectives. For gender identity, this algorithmic reinforcement is not neutral, as it is a deliberate choice of the platform's algorithm. Still, it is very influential and can create confusion, particularly when the gender inferred by social media does not align with a user's self-perception. This misalignment raises questions about how gender identity is constructed and perceived within these platforms.

Transparency, user awareness, and regulatory accountability

Despite the widespread use of GCSs, gender inference remains highly opaque. This opacity operates on multiple levels. First, the platform user experience (UX) design actively conceals information about gender inference. Over time, X has repeatedly modified access pathways (Fosch-Villaronga et al., 2021), making it increasingly difficult for users to locate or modify their inferred gender. On the desktop version, retrieving this information requires navigating multiple menu layers. From the homepage, users must click "More," then "Settings and Privacy," followed by "Your Account," then "Account Information," and finally enter their password before reaching the screen displaying their inferred gender. This five-step, password-protected process suggests an intentional effort to bury gender inference settings deep within the platform's interface. On the mobile app, the situation is even worse, as there appears to be no accessible pathway to view or modify gender inference, effectively preventing users from even knowing how the platform categorizes them.

Beyond UX design, companies are also vague about determining these classifications. While gender inference directly affects users, platforms provide limited or no explanation of the logic behind these decisions. However, transparency is a fundamental principle under EU legislation, particularly the EU AI Act and EU data protection law, especially for the right to rectification, allowing users to correct inaccurate or misleading personal data. Under the General Data Protection Regulation (GDPR), platforms must provide meaningful information about automated decision-making processes. However, gender inference often remains a black box (Bekkum and Borgesius, 2023; Wachter and Mittelstadt, 2019). Without clarity on what signals determine inferred gender or how classifications are assigned, users lack the necessary information to question, rectify, or challenge misclassification—a fundamental requirement for algorithmic accountability. This lack of transparency extends to broader regulatory concerns.

The EU Digital Services Act (DSA) introduces transparency obligations that could improve accountability in these systems. Under Art. 27 DSA, platforms must disclose the main parameters of their algorithmic decision-making, including how gender inference operates and whether users can modify these settings. For Very Large Online Platforms (VLOPs), Art. 38 DSA requires offering at least one non-profiling recommender system, ensuring users can engage with the platform without gender-based categorization affecting content suggestions. Additionally, platforms must report on the potential risks of algorithmic inference, including risks of discrimination, bias, and fairness violations, and what measures are taken to mitigate these harms.

While these obligations represent a regulatory step forward, challenges remain in ensuring that disclosures are accessible and meaningful rather than overly technical or vague at risk that platforms restrict user awareness and limit control over their digital identities.

The risks and consequences of misgendering algorithms

Platforms like X have their reasons for deploying GCSs. These systems can be helpful in further enriching their

datasets that can subsequently be used for predictions, profiling, and automated decision-making (Custers, 2021). Gender is not the only attribute that can be predicted and ascribed; attributes like age, intelligence, sexual orientation, happiness, political orientation, and many more can be predicted, often with high levels of accuracy (Kosinski et al., 2013). Social media platforms use detailed user profiles and characteristics to offer personalized content to individual users and user groups, including customized offers with personalized pricing (Schofield, 2019). While personalized recommendations can enhance user experience, they can also trap individuals in algorithmic filter bubbles, reinforcing pre-existing beliefs and limiting exposure to diverse perspectives (O'Neil, 2016; Pariser, 2011). Users may not even be aware of how their digital identity is shaped by algorithmic classification, increasing their vulnerability to manipulation and disinformation (Plettenberg et al., 2020).

These problems are further exacerbated when accuracy is limited (Custers, 2003). Profiling processes like GCS may amplify bias and inaccuracies and entrench related consequences such as inequalities and discrimination. In case of misgendering, it is not only hard to break out of existing categories and their respective filter bubbles, it may also mean that individuals who are misclassified may receive content, advertisements, or interactions mismatched with their identity, affecting their ability to engage authentically online (Hamidi et al., 2018). Repeated misclassification can lead to emotional distress and identity invalidation, particularly for those who do not conform to binary gender categories (Fergus, 2020; McLemore, 2015). Furthermore, it might harm some of the business cases of online platforms, in case they are less efficient in their targeting.

At the legal level, gender (unlike sexual orientation) is not a special category of data under Art. 9 GDPR, meaning platforms face fewer restrictions in processing it (Bekkum and Borgesius, 2023; Wachter and Mittelstadt, 2019). In other words, gender is not considered a sensitive attribute. Nevertheless, gender is a well-established prohibited discrimination ground across most jurisdictions (Burri and Prechal, 2009). Due to a lack of transparency, discrimination may be complex to detect, and enforcement of anti-discrimination legislation may be complicated. The protection offered by data protection laws may also be complex, given that it is unclear whether inferred data are personal data (Custers and Vrabec, 2024). If so, people have a right to access the inferred gender, including a right to receive meaningful information about the logic involved in data analytics and a right to rectification (Selbst and Powles, 2018). However, invoking these rights requires that people know the GCSs used.

Furthermore, these rights may have their limitations. Typically, the right of access does not include access to the data of others, making it hard to check how inferences were made (Bottis et al., 2019). Invoking the right of rectification may not be ideal for some people, as it requires disclosing additional information about their gender to show how the inferred gender should be rectified (Häuselmann and Custers, 2024). As a result, misgendering is not just a technical failure—it is an issue that raises profound ethical concerns about identity recognition, inclusion, and digital rights.

Methods

Study sample

The data used in the article constitute a cross-sectional novel database (N = 1642) (Fosch-Villaronga et al., 2026). We built the survey using Qualtrics and distributed in English. Participants were recruited via Prolific and had to have a working X account. The answers were collected in November 2023 and processed in 2024. For data quality and rigor, we removed participants who did not provide their Prolific identification number (N = 30), did not complete the survey (N = 52), or had duplicate entries based on the Prolific identification variable (N = 16), leaving us with N = 1553[2] valid responses (S0). The average response time was 5 min and 16 s. The survey's main goal was to explore how individuals perceive gender-based inferences and the challenges of overcoming automated gender assumptions resulting from interface design.

After collecting demographic information —such as age (M = 30.7, SD = 9.9) and continent (Europe: 60.9%, Africa: 27.5%, North America: 4.7%, Asia: 3.8%, Oceania: 1.0%, South America: 1.4%)— we asked participants about four specific aspects related to gender: gender identity (i.e. how one internally defines their gender), biological sex (i.e. the physical characteristics one is born with), sexual orientation, and, finally, gender expression (i.e. the external presentation of gender through actions, clothing, and other behaviors). For gender identity, 48.3% identified as men, 47.1% as women, 3.1% as non-binary, 0.7% as transgender, 0.5% as other, and 0.3% reported no gender. Regarding biological sex, 50.0% were female, 49.7% male, and 0.2% intersex. For sexual orientation, 67.9% identified as straight, 17.6% as bisexual, 7.5% as gay, 2.6% as other, 2.3% as asexual, and 2.3% as questioning. Regarding gender expression, 48.0% reported a masculine expression, 44.7% feminine, 6.0% non-binary, and 1.3% other.

Then, patterns related to both time and user activity on X were collected by analyzing not only the topics users tweeted about (e.g. work-related content, social or political issues) but also how frequently participants used the platform. Next, participants were asked whether they had specified a gender during the registration process or not. At this stage, the sample is divided between those who reported assigning a gender to their profile during registration (S1) and those who did not (S2). The latter group (S2) was used to study the phenomenon of misgendering. To capture how users experience the perception of misclassification,

Table 1. Examples of coded sentiment categories for the question: How do you feel about not being able to remove gender from Twitter?

Sentiment	Responses
Very negative	"I hate it", "I feel angry", "Oppressed", "Very bad", "Outraged", "Helpless and sad", "Horrible", "Very concerned".
Negative	"I think it is wrong", "I do not agree with this we should be able to remove this info", "It is wrong", "It's absurd", "Concerned", "I think it's an unfair thing", "Irritated", "Not happy".
Slightly negative	"Somehow wrong", "Bit annoying", "That's a bit annoying for somebody", "Might be unfair to others", "Slightly bad", "Not so good", "Not very happy".
Neutral	"Neutral", "Don't care", "I don't mind it", "It doesn't bother me", "Not interested", "Not a big deal to me", "No effect", "Indifferent", "No matter", "Not bothered", "Honestly, I don't care"
Slightly positive	"I'm fine with it", "Ok", "Fine", "Accepting", "Fine by me", "it's not a problem for me", "Okey"
Positive	"Good", "Positive", "Great", "I agree", "I like that", "I feel good", "Like it", "Cool"
Very positive "	"It's very good", "Wonderful"

participants were asked to log in to their X account, check the gender that appeared assigned in their profile, and report if, under its consideration, this was right or wrong.

Finally, S1 and S2 were pooled together again to analyze user perceptions. In this sense, user feelings regarding X's gender policies were assessed through two key survey questions: "*A. Should Twitter assume your gender?"* and "*B. How do you feel about not being able to remove gender from Twitter?"* Both were measured using a 7-point Likert scale (Strongly agree – Strongly disagree). Question B was originally collected as an open-ended response and later manually coded into a 7-point Likert scale.[3] To ensure consistency and rigor in the coding process, we developed and applied a coding framework based on clear criteria. Two factors were taken into consideration: firstly, (1) the emotional valence expressed (e.g. negative, neutral, positive), secondly the intensity of language (e.g. use of modifiers like "very" or "somewhat"). Neutral classifications were assigned in cases where evaluative language was absent. Table 1 shows examples of the encoding followed.[4]

Question A was also reformulated as an open-ended response, allowing for a text-based analysis to identify key points that could reveal patterns among users. Thus, to provide a comprehensive overview of users' perceptions, we decided to divide the content of the questions above to analyze participants' feelings toward such practices and the degree of neutrality expressed by users. To this aim, the sample was again divided into three groups: S4, which was used to study levels of neutrality, and S5 (Question A) and S6 (Question B), where neutral responses were excluded to provide a clearer picture of users' feelings. Samples used in this article are illustrated in Figure 1.

Statistical analysis

The aim of our statistical analysis is twofold. The first stage tries to capture the patterns of misgendering along collectives done by the GCS used by X. This first approximation is necessary to analyze later on and understand both the neutrality and perceptions regarding the usage and limits of this technology among users. In this sense, two different types of modelization are used. For the empirical models of the first stage, we used logistic regressions, where our dependent variable takes the value of 1 if the gender that appears in the X account of the participants matches their self-reported gender and 0 otherwise.

However, ordinal logistic regression, precisely the proportional odds (PO) model (McCullagh, 1980: 110), was employed for the second stage, which tended to analyze neutrality and perceptions. The Ordinal Logistic Regression model estimates the effects of a set of independent variables on the logarithm of the probability that the dependent variable assumes low values rather than high values (Eboli and Mazzulla, 2009: 45, Harrell, 2015: 313). This model is what Agresti (2010) calls a cumulative link model. In this sense, the PO model is attractive for dependent variables whose ordinality is based on an underlying continuous dimension, probably unobserved (Eboli and Mazzulla, 2009: 45). This makes it particularly useful for modeling Likert scale outcomes, as in our study, where responses are ordered but not necessarily equidistant (Capuano et al., 2016).

Through this methodological design, we aim to analyze the following hypotheses. Regarding misgendering patterns, we posit that male users experience statistically less misclassification than female ones (H1), straight respondents experience statistically less misclassification than non-straight respondents (H2), and, finally, that users who exhibit a gender expression that does not fit into a binary understanding of gender tend to be more misclassified (H3).

On the other hand, concerning the user's perceptions, we expect that a high level of engagement with the application is associated with decreased neutrality (H4), experiencing misgendering is associated with more negative feelings toward X's gender-related policies (H5), female users are more concerned than male users and show more negative sentiment toward X practices than male users (H6), and, finally, LGBTQ+ users express more concern compared to non-LGBTQ+ users, and show more negative sentiment

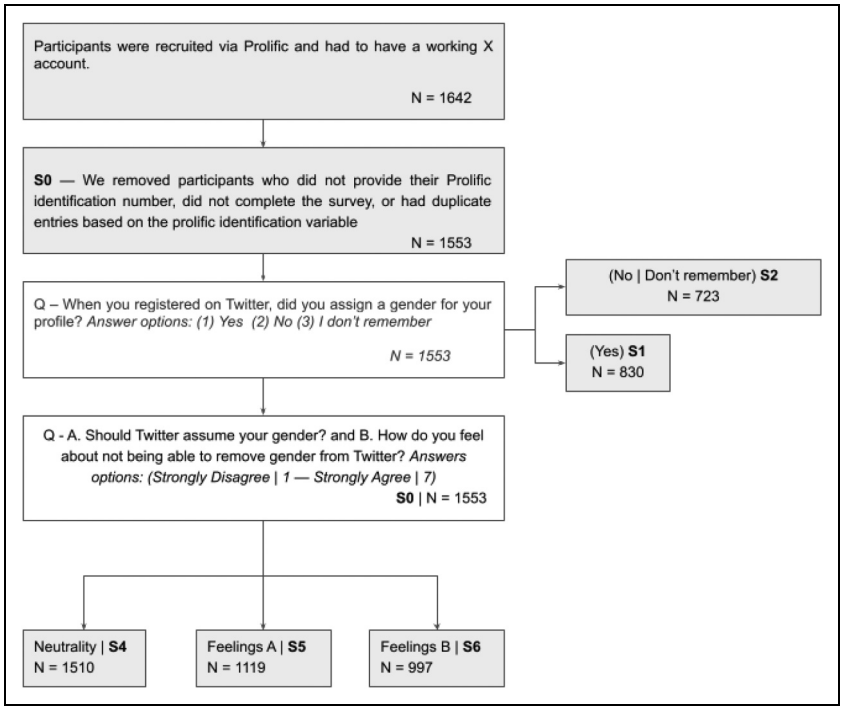

Figure 1. Operationalization of the data used in the analysis.

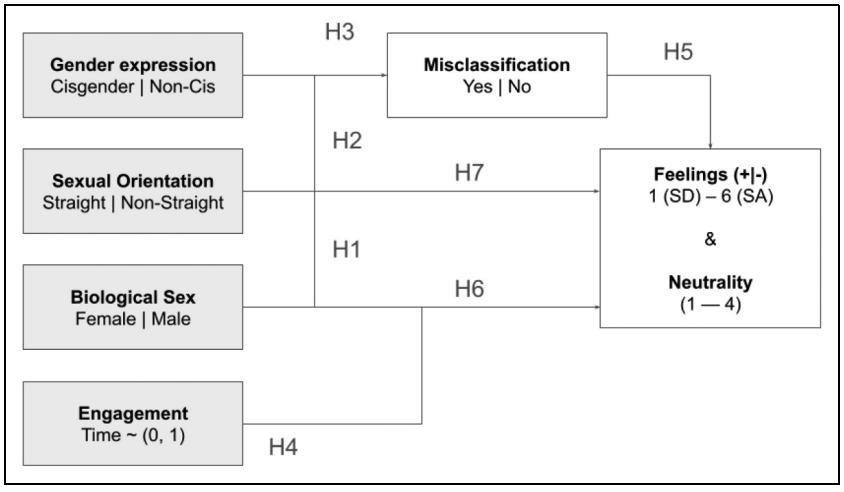

Figure 2. Relations between variables and hypotheses in this study.

toward X practices than non-LGBTQ+ users (H7). The relation between variables and hypotheses is summarized in Figure 2. Table 1 shows the set of regression models used to validate the hypotheses above.

Results

Misgendering algorithms (M1-M4, H1-H3)

Regarding H1, among the 367 women surveyed, 47 (13%) felt that the gender assigned to them was incorrect, compared to 25 men (7%) out of 355. This discrepancy indicates that women are more frequently misclassified than men ($\chi^2 = 6.05$, $p < 0.05$). We re-evaluated misclassification rates to control for potential confounding variables by excluding respondents who identified outside the binary gender framework. Within this refined group, 26 women (8%) out of 339 and 13 men (4%) out of 341 reported misclassification ($\chi^2 = 3.99$, $p < 0.05$), further reinforcing the trend.

On the other hand, to validate H2 and H3, we divided the sample into two groups: respondents identifying as LGBTQ + ($n = 269$) and those who do not ($n = 444$). Among

Table 2. P-value adjustments and effect size for Chi-square tests (H1–H3).

Comparison	χ^2	p-value	Bonferroni Adjusted p	Holm Adjusted p	Significant (Bonferroni)	Significant (Holm)	Cramer's V
Female vs. Male (full sample)	6.05	0.014	0.070	0.028	No	Yes	0.092
Female vs. Male (binary-only)	3.99	0.046	0.230	0.046	No	Yes	0.077
LGTBQ+ vs. Non-LGTBQ+	30.69	<0.01	<0.01	<0.01	Yes	Yes	0.207
Straight vs. Non-straight	28.28	<0.01	<0.01	<0.01	Yes	Yes	0.202
Cisgender vs. Non-Cisgender	91.52	<0.01	<0.01	<0.01	Yes	Yes	0.356

Table 3. Overview of the models used in this study.

Model	Dependent	Independent	Hypothesis	Sample
M1–4	Misgendering (1/0)	Biological sex, sexual orientation and gender expression.	H1, H2 and H3	S2
M5	Feeling A[a] (1–6)	Misgender, Biological sex and Sexual Orientation	H5, H6 and H7	S5
M6	Feeling B[b] (1–6)	Misgender, Biological sex and Sexual Orientation	H5, H6 and H7	S6
M7	Neutrality A (1–4)	Engagement[c], Biological sex and sexual orientation	H4, H6 and H7	S4
M8	Neutrality B (1–4)	Engagement, Biological sex and sexual orientation	H4, H6 and H7	S4

[a] Should Twitter assume your gender?
[b] How do you feel about not being able to remove gender from Twitter?
[c] Engagement refers to the time spent using the application.

non-LGBTQ+ individuals, 429 (92%) reported being correctly classified, while 22 (5%) experienced misclassification. In contrast, 240 (85%) of LGBTQ+ respondents were correctly classified, whereas 50 (15%) reported misclassification. A chi-squared test confirmed a statistically significant difference ($\chi^2 = 30.69$, $p < 0.05$), demonstrating that LGBTQ+ individuals are significantly more likely to be misclassified than their non-LGBTQ+ counterparts. In this sense, a person is classified as part of the LGBTQ+ community if they identify as non-straight or have a gender expression that does not conform to a binary gender framework. Based on these findings, we further examined how misclassification varies across two key factors: sexual orientation and gender expression.

If we analyze H2 and H3 separately, these results remain consistent. Thus, among heterosexual respondents, 433 (95%) out of 457 reported being correctly classified, while only 24 (5%) experienced misclassification. In contrast, among non-heterosexual respondents, 218 (82%) out of 265 were correctly classified, while a significantly more significant proportion, 48 (18%), reported misclassification. A chi-squared test ($\chi^2 = 29.28$, $p < 0.05$) confirmed a statistically significant difference, indicating that non-heterosexual individuals are substantially more likely to be misclassified than their heterosexual counterparts. Moreover, when comparing cisgender ($n = 637$) and non-cisgender ($n = 86$) respondents, 599 (94%) cisgender participants were correctly classified, while 52 (6%) reported misclassification. In contrast, among non-cisgender individuals, only 52 (60%) were correctly classified, whereas 34 (40%) experienced misclassification ($\chi^2 = 91.52$, $p < 0.05$). Table 2 summarizes this first round of statistical tests. We controlled for Type I errors by applying both the Bonferroni and Holm adjustments (Holm, 1979). The majority of comparisons remained statistically significant after correction, reinforcing the robustness of the observed associations. The effect size is calculated through Cramer's V. In this regard, medium to large effects are observed in the comparisons between LGBTQ+ and non-LGBTQ+ individuals, straight and non-straight individuals, and especially between cisgender and non-cisgender individuals.

As discussed, four logistic regression models (M1-M4) were conducted to examine the distribution of misgendering across independent variables (biological sex, gender expression, and sexual orientation). To enhance the robustness of the models, demographics (age and continent), engagement (time spent), and the primary topic driving an individual's X usage were incorporated as control variables. The results from four logistic regression models (M1–M4) are shown in Table 3.

Firstly, M1 examines the impact of biological sex on the likelihood that a respondent affirms their assigned gender. Using female respondents as the reference group, the results indicate that being female significantly decreases the likelihood of reporting alignment with the assigned gender compared to male respondents ($\beta = -0.69$, $p = 0.01$). Secondly,

M2 incorporates gender expression as a predictor, explicitly assessing the effect of non-cisgender expression. The results show a highly significant effect ($\beta = -2.27$, $p < 0.001$), suggesting that individuals who do not conform to cisgender norms are substantially less likely to report alignment with their assigned gender. Thirdly, M3 shifts the focus to sexual orientation, comparing heterosexual and non-heterosexual respondents. The findings indicate that non-straight individuals are significantly less likely to report correct classification than their straight counterparts ($\beta = -1.29$, $p < 0.001$). Finally, M4 integrates all three predictors—biological sex, gender expression, and sexual orientation—into a combined analysis. Gender expression remains the most significant predictor ($\beta = -2.00$, $p < 0.001$), emphasizing its strong influence on alignment between assigned and self-identified gender. However, sexual orientation loses significance in this model ($\beta = -0.48$, $p = 0.17$), suggesting its effect may be mediated by gender expression. Biological sex remains non-significant ($\beta = -0.50$, $p = 0.09$), though it trends toward marginal significance.

Perceptions

Once we detect the systematic misclassification of certain collectives (Section 4.1), our interest delves into analyzing the feelings and perceptions that such practices generate in users. This second analysis stage helps us identify how and why users have a specific position concerning GCSs and, by extension, detect problems that could be addressed through policy and regulatory recommendations (Section 5). To do this, a double approach is followed. Firstly, an exploratory text-based analysis is conducted to identify user concerns and topics of interest. The information extracted from this first stage has also been used to redefine and posit our working hypothesis (H4 to H7). Secondly, the hypotheses mentioned above are studied quantitatively, as explained in Section 3. Table 4 shows the distribution of responses for questions *Should Twitter assume your gender?* and *How do you feel about not being able to remove gender from Twitter?*

Neutrality in responses for both questions as shown in Table 5 is measured using a four-level scale: (1) Strongly Non-Neutral for "Strongly disagree" or "Strongly agree," (2) Moderately Non-Neutral for "Disagree" or "Agree," (3) Slightly Non-Neutral for "Somewhat disagree" or "Somewhat agree," and (4) Fully Neutral for "Neither agree nor disagree." Table 6 shows the distribution for the variable "neutrality" in both questions.

Text-based analysis. Q33 (*How do you feel about Twitter ascribing you a gender female/male based on your Twitter activity?*) is encoded in an open format, enabling respondents to elaborate on their answers. Although these responses have been manually transformed to the aforementioned 7-item Likert scale that later on is used in the quantitative analysis approach (Section 4.2.2), some interesting insights could be extracted from the responses given by X users.

Table 4. Distribution of responses for questions should Twitter assume your gender? And how do you feel about not being able to remove gender from Twitter?

Question	Strongly disagree	2	3	4	5	Strongly agree
A[a]	32%	30%	12%	8%	9%	7%
B[b]	6%	58%	14%	15%	6%	0%

[a] Should Twitter assume your gender?
[b] How do you feel about not being able to remove gender from Twitter?

As indicated, through question 33, users were asked about their feelings regarding X ascribing their gender based on their activity. A significant number of respondents simply indicate their opinion regarding this fact. In this sense, 'Awful,' "Bad," "Wrong," "Terrible," or "Unfair" are among the most common responses within this cluster of respondents. However, other respondents elaborate more on their answers, showing response patterns that could also serve the aim of this article. In this sense, we have identified three core ideas: (1) being misgendered is perceived as a harmful action that could lead to concrete damage to individuals, (2) the legal interest that is perceived as more affected by such practices is privacy, and, (3) a significative number of users declare surprise regarding the commission of these type of practices. Consider that, besides the number of responses about this topic, which could seem small compared to the sample size; the response was neither subject to any restriction nor influenced by any exogenous variables. Thus, the existence of patterns among respondents who freely respond indicates the existence of certain latent traits that could be further explored.

As discussed in Section 2, repeated misclassification can lead to emotional distress and identity invalidation, particularly for those who do not conform to binary gender categories (Fergus, 2020; McLemore, 2015). This is an idea stressed by a significant number of respondents. Thus, comments such as: "I'm not transgender so this doesn't bother me, however I can imagine it would be quite harmful if a trans person was misgendered by twitter," "I don't like it, I'm not really bothered with people/systems calling me a female but some people might feel hurt when the assumption is wrong," "I feel it could hurt someone else who does not feel represented by that gender choice I think this is convenient for twitter however hurtful to its users," constitutes clear examples of the extended idea that besides its legal consequences, misgendering produces tangible and measurable harm upon individuals. This idea supports the necessity to explore the analysis of collectives that are systematically misgendered deeply.

Furthermore, privacy concerns are one of the legal interests that users perceive as more affected by this practice.

Table 5. Distribution of neutrality for questions should Twitter assume your gender? And how do you feel about not being able to remove gender from Twitter?

Question	Strongly Non-Neutral	Moderately Non-Neutral	Slightly Non-Neutral	Fully Neutral
A[a]	27.7%	28.1%	14.7%	24.5%
B[b]	38.9%	31%	17.4%	3.9%

[a] Should Twitter assume your gender?
[b] How do you feel about not being able to remove gender from Twitter?

Table 6. Logistic regression analysis (M1 to M4).

	M1			M2			M3			M4		
Variable	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*	*β*	*SE*	*p-val*
(Intercept)	3.12	1.29	0.0155	3.24	1.36	0.0172	3.90	1.34	0.0036	3.77	1.37	0.0059
Biological Sex (Female)	−0.69	0.28	0.0139	-	-	-	-	-	-	−0.50	0.30	0.0944
Gender Expression (No Cisgender)	-	-	-	−2.27	0.30	<0.0001	-	-	-	−2.00	0.34	<0.0001
Sexual Orientation (No-Straight)	-	-	-	-	-	-	−1.29	0.30	<0.0001	−0.48	0.35	ns
Controls	Yes			Yes			Yes			Yes		
Null deviance:	463.09 on 716 DF			463.09 on 716 DF			463.09 on 716 DF			463.09 on 716 DF		
Residual deviance:	411.16 on 698 DF			363.16 on 698 DF			398.00 on 698 DF			357.88 on 696 DF		
AIC	449.16			401.16			436			399.88		

Notes: ns = p > 0.1.

Comments like "Feel like it would be an intrusion into my privacy," "I feel like my privacy would be invaded," "I want to keep my privacy," "I feel that my privacy is being violated somehow," "Not that big of a deal for me in terms of gender identity but mostly a privacy problem," "Insecure about my privacy," "I feel it is a detriment to my privacy," reflects that idea essentially. Related to this, profiling is another problem that users stress. In this sense, it is possible to observe responses like "Seems like profiling" or "I don't like being profiled by social media."

Surprise among respondents, in a certain way, confirms the outlined opacity of these kinds of practices (Section 2.3). An important number of users declare being surprised by the existence of gender assignment. Illustrative examples of these could be found in commentaries like: "Odd," "Strange," "Would be kinda weird and unnecessary," "Strange concept," or "It feels weird because I have not mentioned my gender prior." Thus, although GCSs are widely used across digital platforms, users demonstrate ignorance when reading about the usage of these practices. Indeed, the absence of knowledge regarding the usage of GCSs and the technology in itself ("Weird, I don't know how that works") could lead to a broader misunderstanding or to making presumptions not based on evidence regarding how this technology works. See, for instance, the commentary: "I find it a bit weird that twitter assigned me a male; I almost feel like it was sexist because I post about *research/politics mainly* - is twitter implying that these are male activities and not female?". So, despite transparency duties (Section 2.3), users remain uninformed about such practices, which underlines the inefficiency of transparency-based interventions.

Ordinal logistic regression analyses. Table 7 summarizes the final ordinal logistic regression models for feelings (M5 and M6) and neutrality (M7 and M8). The models were estimated using proportional odds logistic regression using R (version 4.4.2). Logit was used as a link function.

Age and topics have been used as controls for all models. For all the analyses, the topics for which X is used were not significant. Age was not significant for models M5 and M6, but shows a slight negative correlation with neutrality for ages between 26 and 37. For the remaining models and age categories, the odd proportional tests were not significant, suggesting that the effects were proportional across the categories of the outcome variable.

Regarding H4, time spent using the application is not a significant predictor in M5 and M6, suggesting no meaningful impact on responses in these models. However, consistent with our hypothesis (H4), in M7 (OR = 0.83), individuals who engaged with the system for extended periods were 17% less likely to select a higher response category, indicating a negative association with neutrality or agreement. This trend persisted in M8 (OR = 0.87), though the effect was slightly weaker, with a 13% lower likelihood of expressing neutrality or favorable agreement.

Furthermore, being misgendered is a significant predictor in most models, with affected individuals consistently

Table 7. Summary of ordinal logistic regression analysis.

Dependent Variable	M5 (Feelings)			M6 (Feelings)			M7 (Neutrality)			M8 (Neutrality)		
	Gender Inference			Lack of option			Gender Inference			Lack of option		
	β	SE	p-val	β	SE	p-val	β	SE	p-val	β	SE	p-val
Biological Sex (Female)	−0.63	0.11	<0.001	−0.04	0.13	ns	−0.46	0.1	<0.0001	−0.30	0.1	0.0027
Sexual Orientation (Not Straight)	−0.23	0.12	<0.055	−0.53	0.14	0.00015	−0.2	0.10	0.045	−0.53	0.11	<0.0001
Misgendered	−0.47	0.22	<0.0032	−1.13	0.27	<0.0001	−0.32	0.2	ns	−0.50	0.22	0.0023
Age[a] [23, 26]	−0.17	0.16	ns	0.17	0.18	ns	−0.17	0.14	ns	−0.2	0.14	ns
Age [27, 30]	−0.23	0.16	ns	−0.08	0.19	ns	−0.27	0.14	0.054	−0.39	0.15	0.0094
Age [31, 37]	−0.1	0.17	ns	−0.05	0.20	ns	−0.25	0.15	0.095	−0.16	0.15	ns
Age [38, 79]	0.27	0.17	ns	0.13	0.20	ns	−0.10	0.15	ns	−0.16	0.15	ns
Actuality (i.e. News, politics)	0.18	0.39	ns	−0.42	0.43	ns	0.1	0.33	ns	−0.37	0.33	ns
Social	−0.12	0.39	ns	−0.37	0.43	ns	0.16	0.32	ns	−0.27	0.33	ns
Time	−0.01	0.06	ns	0.06	0.06	ns	−0.19	0.05	0.00015	−0.14	0.1	ns

Notes: Reference category has been omitted for simplicity. ns = $p > 0.1$.

[a] The age distribution was designed using a dual criterion: first, to achieve a balanced distribution across the sample, and second, to incorporate a generational perspective that enables us to control for potential hidden patterns that may be generationally driven. Based on this approach, we established the following age groups: (1) 23–26: This group is composed of Generation Z, (1) 27–30: A transitional group (late Gen Z and early Millennials) capturing individuals in a key developmental shift from early adulthood to more stable adult roles, (3) 31–37: This range includes Millennials in their prime adult years, often associated with career consolidation, family formation, and increased societal engagement, and (4) 38–75: A broader group encompassing Generation X and Baby Boomers.

reporting lower agreement or neutrality compared to those who were not misgendered (H5). Specifically, in M5 (OR = 0.63), misgendered individuals had 37% lower odds of expressing positive or neutral feelings. The effect was strongest in M6 (OR = 0.32), where their odds were 68% lower, indicating a substantial negative association regarding the algorithm's control over gender inference. Also, in M8 (OR = 0.61), misgendered individuals had 39% lower odds of reporting neutrality or positive agreement regarding the lack of an option to change their inferred gender.

Biological sex was a significant predictor in all analyses except M6. Specifically, female respondents were more likely to report either negative feelings (M5: OR = 0.53) or lower levels of neutrality (M7: OR = 0.63; M8: OR = 0.74) compared to male respondents (H6). In this sense, for female respondents, the odds of agreeing with the algorithm's inference of gender are 47% lower than for male respondents and approximately 38% lower for being neutral. Regarding the lack of control over the outcome, female respondents have 27% lower odds of holding a neutral opinion than male respondents.

Finally, sexual orientation was also a significant predictor across all models (H7), with non-straight individuals consistently showing lower odds of positive or neutral responses compared to straight individuals. Individuals who do not identify as straight were consistently less likely to report positive or neutral feelings than their straight counterparts (M5: OR = 0.79; M6: OR = 0.59; M7: OR = 0.82; M7: OR = 0.59). Specifically, non-straight individuals had 21% lower odds of expressing a positive attitude toward gender inference. Regarding the algorithm's control over outcomes, their odds of a higher response category were 41% lower, indicating a strong negative association. While the effect was smaller but still notable for neutrality (18% lower odds), the reduction remained substantial in M8, where non-straight individuals were 41% less likely to report neutrality or positive agreement regarding the lack of an option to change the inferred gender.

Conclusions

Our findings indicated that certain marginalized groups, namely women and the LGBTQ+ community, are systematically misclassified and therefore misgendered by GCSs (Fosch-Villaronga et al., 2026). Disproportionate inaccuracies affecting these historically discriminated groups highlight a persistent data gap (Sperber et al., 2023) in information about aspects of women's and gender and sexually diverse people's lives and bias in favor of men and straight people, exacerbating existing prejudice.

Analysis of user responses collected in this study reveals that gender misclassification is in and of itself perceived as a significant harm for those who experience it. Furthermore, gender misclassification has a perceived knock-on effect, risking subsequent emotional distress and identity invalidation, which is especially detrimental to gender and sexually diverse individuals (Fergus, 2020; McLemore, 2015). Our analysis shows that while some respondents reported indifference to misclassification, others acknowledged its potential harm, particularly for transgender individuals. This is particularly concerning for gender-diverse people who often turn to social media due to a lack of traditional support for self-preservation, improved physical safety, protection from harassment, and community building (Katyal and

Jung, 2021) as it undermines perceived psychological safety online. These findings align with broader discussions on psychological privacy harms (Citron and Solove, 2022; Kilovaty, 2021), as misclassification can lead to distress, a diminished sense of identity, and exclusion. Thus, ensuring algorithmic accuracy is not merely a technical challenge that needs to be confronted and a legal requirement outlined in the EU AI Act, but a fundamental issue of safety, fairness, and human dignity.

Beyond emotional and psychological harms, gender classification systems (GCSs) also pose significant privacy and profiling risks that extend into the legal and regulatory domain. Under regulations like the GDPR, inferred gender paradoxically isn't classified as "sensitive" data (unlike race or sexual orientation) and thus lacks special protection. However, users in our study perceived gender inferences as deeply personal and identified privacy concerns as one of the legal interests most affected by automated gender classification. Many respondents voiced a loss of agency and unease with these hidden data collection practices. This sentiment applies even in those cases where those practices lead to more accurate outcomes, as echoed by research showing that highly accurate algorithmic profiles can leave individuals feeling surveilled and powerless (Zhang et al., 2025). Respondents furthermore questioned the necessity and utility of gender inference for service delivery, suspecting it serves mainly advertising or profiling purposes. Indeed, even legal scholars note that using "legitimate interest" as a basis for inferring gender is contentious and not justified, underlining the lack of any evidence that such gender profiling is truly necessary (Fosch-Villaronga et al., 2020). These findings align with broader critiques of automated decision-making systems, which warn that opaque profiling practices can reinforce gender binarism, undermine autonomy, surveil individuals, and threaten safety (Hamidi et al., 2018).

The opaque design of social media interfaces further compounds these legal and ethical challenges, often concealing how gender inferences are made and where they are stored. In this sense, our results confirm the opacity of social media UX design when it comes to inferred gender: users find it challenging to locate where such information is displayed or to change it. Findings underscore this challenge—many participants are unaware of their inferred gender until prompted, and they struggle to navigate the labyrinthine settings to modify it. This surprise indicates that current transparency mechanisms are inadequate; people simply do not know how these inferences are being made. Even when disclosed, the controls are limited: users can edit their assumed gender in some cases, but they cannot entirely prevent the profiling without opting out of broad personalization features (Fosch-Villaronga et al., 2020). Such design choices keep users in the dark and hinder autonomy.

In light of this opacity, calls for improved transparency and explainability have emerged as potential solutions to restore user autonomy and address information asymmetry. Clear disclosure and explainability around the working of GCSs about the gender inferences they produce can help bridge the information asymmetry between companies and users, putting individuals in a better position to understand and contest how they are categorized (Lütz, 2024). In theory, this improved transparency could empower some users to correct misclassifications or demand opt-out options—a reaction supported by prior work showing that people who feel surveilled by accurate profiling are more inclined to adjust their privacy settings or ad preferences (Zhang et al., 2025). However, our work shows that simply informing users that a binary gender has been assigned offers little comfort or accuracy for those who do not fit the binary model, as many systems still only offer and distinguish between "male" or "female" categories. While laws like the GDPR and the EU AI Act mandate transparency to ensure users are aware of their interaction with an AI system and give them rights to access or rectify their data, in practice, users face notable hurdles in asserting control over their gender representation. Notably, the EU AI Act requires explainability measures to be implemented into the design and development of an AI system in such a way as to ensure deployers can easily understand its operation, interpret its output, and use it appropriately (article 13). However, this requirement only applies to so-called "high-risk" AI systems (article 6) and their deployers (Article 3(4)), both of which do not apply to the users of X in any case. Despite mandated accountability for social media platforms engaging in automated gender classification practices, our findings suggest a disconnect between these formal requirements and the reality of an opaque, binary-focused system that limits users' practical ability to manage their own digital identity.

Moreover, our results show that user perceptions regarding gendering practices depend heavily on whether people are gendered correctly or incorrectly. People who are misgendered find these practices problematic, perceiving them as a form of discrimination, and feel vulnerable. However, people who are gendered correctly report indifference (in their own words, they are mostly 'fine' with this practice). In the end, the majority of the people either provide their gender themselves or are gendered correctly, which means the group of misgendered people is relatively small. However, the groups of people that are at risk of being misgendered are exactly the groups that already are generally considered more marginalized, such as women compared to men, non-straight persons compared to straight persons, and non-cisgender persons compared to cisgender persons. These vulnerabilities may be exacerbated because of intersectionality (Weldon, 2008), in which intersecting and overlapping social identities may amplify issues, such as privacy issues (e.g. predicting gender that people may not want to disclose) and discrimination issues (e.g. using predicted gender as a basis for decision-making). In general,

GCSs may violate aspects of human dignity, as misgendering may affect perceptions people have regarding their identity, autonomy, self-esteem, and personal development. From this perspective, misgendering practices have a much broader impact than only discriminatory aspects (Orwat, 2024).

Despite these significant harms, the relatively small size of the affected population means that platforms and policymakers may lack sufficient incentives to enact meaningful change. Platforms and regulators/legislators are often influenced by the average consumer's or voter's views. In democratic societies, it is essential to focus on majorities and consider the needs and preferences of minorities, particularly vulnerable minorities (Wheatley, 2005). Also, from a moral perspective, this is important. Any moral perspective has as a starting point that the vulnerabilities of others (or, more in general, the vulnerability of human beings and the human condition) are taken into account (Frankena, 1973). From these perspectives, it can be argued that both platforms and governments have a moral responsibility to regulate gendering practices in such a way that they prevent misgendering as much as possible (which might include refraining from these practices or prohibiting them) and to protect vulnerable groups for which these practices are problematic and harmful.

In light of these findings, transparency alone may be insufficient to legitimize these practices. While regulators may focus on ensuring greater transparency and user control, our results suggest that user perceptions should be more critical in guiding regulatory interventions. Misgendering not only reinforces discriminatory outcomes but also risks deepening social exclusion for already vulnerable groups. Consequently, policymakers should reassess the legality of gender inference in certain contexts under privacy and anti-discrimination laws. This could shift the debate toward stricter limitations on these practices—such as implementing opt-out mechanisms, requiring explicit user consent, or even prohibiting gender classification in specific contexts—better to protect user autonomy, dignity, and inclusion. That said, we acknowledge the practical limits of influencing platform behavior—particularly on X, where recent shifts reflect growing hostility toward gender diversity. While direct change may be unlikely, documenting misgendering remains essential for informing public debate, legal scrutiny, and future platform design elsewhere.

In conclusion, this study has demonstrated that algorithmic gender classification on X frequently misgenders marginalized and minority population groups, with errors disproportionately affecting women and LGBTQ+ individuals and those who do not fit within the gender binary. Our findings reveal that users are not neutral concerning misgendering awareness; those who experienced misgendering reported heightened aversion toward the platform's gender inference practices, often expressing frustration, harm, and a loss of agency. Knowing they are misgendered leads to aversion towards gender classification systems. Importantly, our work illustrates that misgendering is not merely perceived as a regulatory oversight and technical failure but as a violation of individuals' freedom to define and control their identity. These insights expose the power asymmetry between users and social media platforms, where individuals are subjected to opaque gender inference practices with minimal control or awareness.

Acknowledgments

This article is part of the ERC StG Safe and Sound project, a project that has received funding from the European Union's Horizon-ERC program, Grant Agreement No. 101076929. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

ORCID iDs

Eduard Fosch-Villaronga https://orcid.org/0000-0002-8325-5871
Antoni Mut-Piña https://orcid.org/0000-0001-7841-5992
Tessa Verhoef https://orcid.org/0000-0002-1219-3730
Adam Poulsen https://orcid.org/0000-0002-0001-3894
Roger A. Søraa https://orcid.org/0000-0001-6800-0558
Hadassah Drukarch https://orcid.org/0000-0001-9695-8990
Anne Noel https://orcid.org/0009-0009-9066-6407
Scarlet Tunney https://orcid.org/0009-0002-5896-0760
Bart Custers https://orcid.org/0000-0002-3355-8380

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the H2020 European Research Council Starting Grant Safe & Sound (grant number 101076929).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Data will be made available on request.

Notes

1. Lesbian, gay, bisexual, transgender, gender diverse, intersex, queer, asexual and questioning persons.
2. Please note that the three exclusion criteria are not mutually exclusive. This is why, if they are subtracted to the original sample size, the result is not the same. In this case, 28 respondents neither provided their prolific identification number nor finished the survey.
3. Three researchers converted the variable into a 7-point scale. The open-ended questions were converted into a 7-point scale, where the categories represent feelings towards 'not being able

to remove gender from Twitter'. While we did not calculate a formal intercoder reliability statistic, several steps were taken to ensure consistency and rigor in the coding process. We developed and applied a coding framework based on several clear criteria, including the emotional valence expressed (e.g. negative, neutral, positive), the intensity of language (e.g. use of modifiers like "very" or "somewhat"), and the overall tone. Neutral classifications were assigned in cases where evaluative language was absent or unclear.

4. Raw data is available from the authors upon request.

References

Agresti A (2010) Modeling ordinal categorical data. *Anal Ordinal Categ Data* 75(10.1002): 9780470594001.

Alvarado Garcia A, Yang T and Miceli M (2025) What knowledge do we produce from social media data and how?*Proceedings of the ACM on Human-Computer Interaction* 9(1): 1–45.

Al Zamal F, Liu W and Ruths D (2012) Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In: *Proceedings of the international conference on weblogs and social media*.

Argamon S, Koppel M, Pennebaker JW, et al. (2007) Mining the blogosphere: age, gender and the varieties of self-expression. *First Monday 12* 9(3).

Barlas P, Kyriakou K, Guest O, et al. (2021) To "see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW3): 1–31.

Bekkum M and Borgesius FZ (2023) Using sensitive data to prevent discrimination by artificial intelligence: does the GDPR need a new exception? *Computer Law & Security Review* 48: 105770.

Bijker WE (1994) Sociohistorical technology studies. In: *Handbook of Science and Technology Studies*. Thousand Oaks, CA: SAGE, 229–256.

Bivens R and Haimson OL (2016) Baking gender into social media design: how platforms shape categories for users and advertisers. *Social Media + Society* 2(4). DOI: 10.1177/2056305116672486.

Bottis M, Panagopoulou-Koutnatzi F, Michailaki A, et al. (2019) The right to access information under the GDPR. *International Journal of Technology Policy and Law* 3(2): 131–142.

Bray F (2007) Gender and technology. *Annual Review of Anthropology* 36(1): 37–53.

Burger JD, Henderson J, Kim G, et al. (2011, July) Discriminating gender on Twitter. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp.1301–1309.

Burri S and Prechal S (2009) Comparative approaches to gender equality and non-discrimination within Europe. In: *European Union Non-Discrimination Law*. London: Routledge-Cavendish, 247–280.

Capuano AW, Dawson JD, Ramirez MR, et al. (2016) Modeling Likert scale outcomes with trend-proportional odds with and without cluster data. *Methodology* 12(2): 33–43.

Chen X, Wang Y, Agichtein E, et al. (2015) A comparative study of demographic attribute inference in twitter. *Proceedings of the International AAAI Conference on Web and Social media* 9(1): 590–593.

Cinelli M, De Francisci Morales G, Galeazzi A, et al. (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118(9): e2023301118.

Citron DK and Solove DJ (2022) Privacy harms. *Boston University Law Review* 102: 793.

Cockburn C and Ormrod S (1993) *Gender and Technology in the Making*. London, UK: Sage Publications Ltd.

Costanza-Chock S (2020) *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

Criado-Perez C (2019) *Invisible Women: Exposing Data Bias in a World Designed for Men*. London: Vintage.

Custers BHM (2003) Effects of unreliable group profiling by means of data mining. In: Grieser G, Tanaka Y and Yamamoto A (eds) Lecture notes in artificial intelligence, proceedings of the 6th international conference on discovery science (DS 2003), Vol. 2843, Sapporo, Japan, pp.290–295. Berlin, Heidelberg, New York: Springer-Verlag.

Custers BHM (2021) Profiling and predictions: challenges in cybercrime research datafication. In: Lavorgna A and Holt T (eds) *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*. Cham, Switzerland: Palgrave MacMillan, pp.63–79.

Custers BHM and Vrabec H (2024) Tell me something new: data subject rights applied to inferred data and profiles. *Computer Law & Security Review* 52: 105956.

Eboli L and Mazzulla G (2009) A new customer satisfaction index for evaluating transit service quality. *Journal of Public Transportation* 12(3): 21–37.

Ellison NB, Pyle C and Vitak J (2022) Scholarship on well-being and social media: a sociotechnical perspective. *Current Opinion in Psychology* 46: 101340.

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

Fergus J (2020) Twitter is guessing users' genders to sell ads and often getting it wrong. *Input*. Available at: https://www.inputmag.com/tech/twitter-guesses-your-gender-to-serve-you-ads-relevant-tweets-wrong-misgendered (last modified 28 April 2020).

Fosch-Villaronga E and Malgieri G (2024) Queering the ethics of AI. In: Gunkel DJ (ed) *Handbook on the Ethics of Artificial Intelligence*. Northampton, MA: Edward Elgar Publishing, 301–315.

Fosch-Villaronga E, Mut-Piña A, Verhoef T, et al. (2026) Misgendering algorithms: insights from a cross-sectional survey on algorithmic gender classification in social media. *Technology in Society* 86(103110): 1–15.

Fosch-Villaronga E, Poulsen A, Søraa RA, et al. (2020) Don't guess my gender, gurl: the inadvertent impact of gender inferences. In: Bias and fairness in AI: workshop at ECMLPKDD 2020, Ghent, Belgium, 18 September 2020, pp.1–9: Springer.

Fosch-Villaronga E, Poulsen A, Søraa RA, et al. (2021) A little bird told me your gender: gender inferences in social media. *Information Processing & Management* 58(3): 102541.

Frankena WK (1973) *Ethics*. Englewood Cliffs, NJ: Prentice Hall.

Garimella K, De Francisci Morales G, Gionis A, et al. (2018, April) Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. In: *Proceedings of the 2018 World Wide Web conference*, pp.913–922.

Hamidi F, Scheuerman MK and Branham SM (2018, April) 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.

Harrell Jr FE (2015) *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer International Publishing.

Häuselmann A and Custers B (2024) The right to rectification and inferred personal data. *European Journal of Law and Technology* 15(3): 1–24.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2): 65–70. JSTOR 4615733. MR 0538597.

Johnston SF (2018) The technological fix as social cure-all: origins and implications. *IEEE Technology and Society Magazine* 37(1): 47–54.

Karkazis K (2019) The misuses of "biological sex". *The Lancet* 394(10212): 1898–1899.

Katyal SK and Jung JY (2021) The gender panopticon: AI, gender, and design justice. *UCLA Law Review* 68(3): 692–785.

Keyes O (2018) The misgendering machines: trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interactionn* 2: 1–22.

Kilovaty I (2021) Psychological data breach harms. *North Carolina Journal of Law & Technology* 23: 1.

Kosinski M, Stillwell D and 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.

Liu W and Ruths D (2013) What's in a name? Using first names as features for gender inference in Twitter. In: *Analyzing Microtext*: 2013 AAAI spring symposium.

Lorenz TK (2021) Relying on an "other" category leads to significant misclassification of sexual minority participants. *LGBT health* 8(5): 372–377.

Lütz F (2024) The AI Act, gender equality and non-discrimination: what role for the AI office? *ERA Forum* 25(1): 79–95. Berlin/Heidelberg: Springer Berlin Heidelberg.

McCullagh P (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42(2): 109–127.

McLemore KA (2015) Experiences with misgendering: identity misclassification of transgender spectrum individuals. *Self and Identity* 14(1): 51–74.

Mueller J and Stumme G (2016, August) Gender inference using statistical name characteristics in Twitter. In: *Proceedings of the the 3rd multidisciplinary international social networks conference on socialinformatics 2016, data science 2016*, pp.1–8.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Orwat C (2024) Algorithmic discrimination from the perspective of human dignity. *Social Inclusion* 12: 1–18.

Pariser E (2011) *The Filter Bubble: What the Internet Is Hiding from You*. London: Penguin.

Pino M and Edmonds DM (2024) Misgendering, cisgenderism and the reproduction of the gender order in social interaction. *Sociology* 58(6): 1243–1262.

Plettenberg N, Nakayama J, Belavadi P, et al. (2020). User behavior and awareness of filter bubbles in social media. In: *Digital human modeling and applications in health, safety, ergonomics and risk management. 11th international conference, DHM 2020*, Copenhagen, Denmark, 19–24 July 2020, Proceedings, Part II 22, pp.81–92: Springer.

Poulsen A, Fosch-Villaronga E and Søraa RA (2020) Queering machines. *Nature Machine Intelligence* 2(3): 152–152.

Rentetzi M (ed) (2023) *The Gender of Things: How Epistemic and Technological Objects Become Gendered*. New York, NY: Taylor & Francis.

Ruberg B and Ruelos S (2020) Data for queer lives: how LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society* 7(1): 205395172093328.

Ruppert E (2018) *Sociotechnical Imaginaries of Different Data Futures*. Rotterdam, Netherlands: Erasmus University Rotterdam.

Sakaki S, Miura Y, Ma X, et al. (2014) Twitter user gender inference using combined analysis of text and image processing. In: *Proceedings of the third workshop on vision and language*, Dublin, Ireland, pp.54–61: Dublin City University and the Association for Computational Linguistics.

Schofield A (2019) Personalized pricing in the digital era. *Competition Law Journal* 18(1): 35–44.

Schroeder J (2021) Reinscribing gender: social media, algorithms, bias. *Journal of Marketing Management* 37(3-4): 376–378.

Selbst A and Powles J (2018) "Meaningful information" and the right to explanation. In: *Conference on fairness, accountability and transparency*, pp.48–48: PMLR.

Søraa RA and Bruijning N (2024) Gendering the boundary object: "Sophia the robot" as cyborg-woman, fashionista, citizen, and imagination. In: *The Gender of Things*. New York, NY: Routledge, 152–166.

Sperber S, Täuber S, Post C, et al. (2023) Gender data gap and its impact on management science — reflections from a European perspective. *European Management Journal* 41(1): 2–8.

Ur B, Leon PG, Cranor LF, et al. (2012, July) Smart, useful, scary, creepy: perceptions of online behavioral advertising. In: Proceedings of the eighth symposium on usable privacy and security, pp.1–15.

Vásárhelyi O and Brooke S (2025) Computing gender. In: Yasseri T (ed) *Handbook of Computational Social Science*. Cheltenham, UK: Edward Elgar Publishing.

Verbeek PP (2012) Expanding mediation theory. *Foundations of Science* 17(4): 391–395.

Vivienne S, Hanckel B, Byron P, et al. (2023) The social life of data: strategies for categorizing fluid and multiple genders. *Journal of Gender Studies* 32(5): 498–513.

Wachter S and 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.

Weldon SL (2008) Intersectionality. In: *Politics, Gender and Concepts: Theory and Methodology*. Cambridge, UK: Cambridge University Press, 193–218.

Wheatley S (2005) *Democracy, Minorities and International Law*. Cambridge, UK: Cambridge University Press.

Wiegand AA, Sheikh T, Zannath F, et al. (2024) "It's probably an STI because you're gay": a qualitative study of diagnostic error experiences in sexual and gender minority individuals. *BMJ Quality & Safety* 33(7): 432–441.

Zhang D, Strycharz J, Boerman SC, et al. (2025) Google knows me too well! Coping with perceived surveillance in an algorithmic profiling context. *Computers in Human Behavior* 165: 108536.