Journal of Computer-Mediated Communication, 2024, zmad053
https://doi.org/10.1093/jcmc/zmad053
Full-length Research Article
Special Issue: Gender Gaps in Digital Spaces

Categorizing the non-categorical: the challenges of studying gendered phenomena online

Sarah Shugars [1,*], Alexi Quintana-Mathé [2], Robin Lange[2], David Lazer[2,3,4,5]

[1]Department of Communication, Rutgers University, New Brunswick, NJ, USA
[2]Network Science Institute, Northeastern University, Boston, MA, USA
[3]Department of Political Science, Northeastern University, Boston, MA, USA
[4]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[5]Institute of Quantitative Social Science, Harvard University, Cambridge, MA, USA
*Corresponding author: Sarah Shugars. Email: sarah.shugars@rutgers.edu

Abstract

Studies of gendered phenomena online have highlighted important disparities, such as who is likely to be elevated as an expert or face gender-based harassment. This research, however, typically relies upon inferring user gender—an act that perpetuates notions of gender as an easily observable, binary construct. Motivated by work in gender and queer studies, we therefore compare common approaches to gender inference in the context of online settings. We demonstrate that gender inference can have downstream consequences when studying gender inequities and find that nonbinary users are consistently likely to be misgendered or overlooked in analysis. In bringing a theoretical focus to this common methodological task, our contribution is in problematizing common measures of gender, encouraging researchers to think critically about what these constructs can and cannot capture, and calling for more research explicitly focused on gendered experiences beyond a binary.

Lay Summary

A growing body of work has highlighted the challenge of online gender gaps—for example, women typically receive fewer likes, follows, and shares than men. However, efforts to measure these gaps are complicated by users' ability to customize the gendered signals shared through their name, biography, photo, or other aspects of their online presence. Researchers typically rely on these visible signals to estimate users' gender. Yet, such a process might misclassify or discount those who do not make their gender visible or whose gender does not conform to a binary, cisgender norm. To examine this challenge, we compare three measures of users' gender: (1) administrative records, (2) use of pronouns or other gendered terms, and (3) human hand-coding. We investigate variations in these measures and examine impacts on gender-based estimates of attention and amplification. Consistent with past research, we find that women tend to receive less attention than men, though these gaps may be sensitive to the measure of gender used. Furthermore, we find that nonbinary users and those without a visible gender may face unique challenges in online spaces and typically receive less attention than their binary peers.

Keywords: gender, affordances, attention, amplification, gender disparities.

Since the early days of the Internet, researchers in computer-mediated communication have conducted critical research aimed at identifying and ameliorating gendered inequities in online participation and experiences (Kapidzic & Herring, 2011; Nilizadeh et al., 2021; Savicki et al., 1996; Witmer & Katzman, 1997). Such work has repeatedly illustrated troubling gender gaps in attention and amplification (Messias et al., 2017; Nilizadeh et al., 2021; Shugars et al., 2021), as well as gendered patterns in online harassment (Döring & Mohseni, 2019; Elias & Gurbanova, 2018; Megarry, 2014; Nadim & Fladmoe, 2021). A key challenge to this work, however, is the performative nature of gender (Butler, 1990) and how that performance plays out in the context of a platform affordance (Evans et al., 2017; Treem & Leonardi, 2013) we call "gender visibility."

While interpersonal interactions in all contexts can be understood as a performance of the self (Goffman, 1959), online settings afford users unique opportunities to easily customize and control many aspects of their self-presentation. Particularly relevant to the performance of gender, online users can typically select an avatar to represent themselves, create an original name, and choose whether or not to share pronouns or other gendered terms in their bios. In other words, the affordance of gender visibility means that users can not only customize how their gender is expressed but can choose to represent themselves as a different gender than they typically perform in their daily lives (Armentor-Cota, 2011; Lehdonvirta et al., 2012; Ratan et al., 2019).

The affordance of gender visibility then raises difficult questions about how researchers should interact with the construct of gender when investigating gendered phenomena online. For example, researchers might reasonably want to study variations in how much attention people of different genders typically receive. Previous work has found that men—or, more precisely, accounts researchers identified as belonging to men—regularly receive more interaction with and amplification of their content (Messias et al., 2017; Nilizadeh et al., 2021; Shugars et al., 2021). This important finding speaks to gender disparities in who is accepted as an "expert" and can have meaningful offline consequences (Gallagher et al., 2021; Jackson et al., 2020). However, measuring gender-based disparities requires estimating users' gender—a challenge since users perform gender both on and offline (Butler, 1990; Kapidzic & Herring, 2011; Lehdonvirta et al., 2012; Morgan et al., 2020; Savicki et al., 1996; Witmer & Katzman, 1997).

Associate Editors: Cuihua Shen and Emőke-Ágnes Horvát
Received: 5 April 2023. **Revised:** 22 November 2023. **Accepted:** 4 December 2023

This is the tension that motivates this article. We believe strongly that it is important for researchers to document and work to minimize gendered gaps in online settings. Yet, too often, the challenges inherent in this work are minimized as merely methodological limitations. As a research team led by a nonbinary first author and with half our team transgender or nonbinary, we are particularly concerned about the tendency for research to assume a binary gender classification of "Male" or "Female" and to drop users who do not fit neatly into the expectations for these categories. While this is a common approach in the literature, doing so definitionally misclassifies nonbinary users and actively erases the experiences of anyone who does not perform their gender in narrowly defined ways. Such work may then unintentionally contribute to increasing gender gaps by perpetuating notions of gender as an easily observable, binary construct.

In this article, we therefore draw on work in gender and queer studies to complicate the examination of gendered phenomena in online settings. We focus specifically on three approaches that are commonly used in the literature to infer the gender of online users: administrative records (Pagnucci & Mauriello, 1999; Shugars et al., 2021), linguistic signals (Jiang et al., 2022; Kapidzic & Herring, 2011; Nilizadeh et al., 2021; Tucker & Jones, 2023), and hand-coding of accounts (Döring & Mohseni, 2019; Elias & Gurbanova, 2018; Savicki et al., 1996; Witmer & Katzman, 1997). We explore these measures' theoretical and methodological strengths and weaknesses, showing that gender inference methods can have downstream consequences for interpreting gendered phenomena. Using a dataset of 1.6 million Twitter users matched to U.S. voting records (Grinberg et al., 2019; Shugars et al., 2021), we focus specifically on how different estimates of user gender can influence understandings of gender gaps in online attention and amplification.

There is no perfect way to capture gender; a complex construct that may change over time and be subject to varied interpretations (Billard, 2019; Butler, 1990, 2004; Guyan, 2022). Our goal in this article is not to present a singular way of measuring gender or to indict the incredible scholars who have used these approaches to advance our understanding of online gender dynamics. Rather, in explicitly bringing a theoretical focus to this common methodological task, our contribution is in problematizing common measures of gender, encouraging researchers to think critically about what these constructs can and cannot capture, and calling for more research explicitly focused on the gendered experiences of users beyond the gender binary or who perform their gender in ways outside a cisgender norm.

Gender and self-presentation

Goffman's dramaturgical theory famously argues that all interpersonal interaction involves a presentation of the "self" (Goffman, 1955, 1959). People perform differently for their parents than for their friends; they choose what to say, how to say it, and select aspects of themselves to make salient in different contexts. Butler's seminal work in gender studies similarly turns to the idea of "performance" and interprets identity as an action rather than an innate trait (Butler, 1990). As Butler writes, " … acts of gender create the idea of gender, and without those acts, there would be no gender at all" (Butler, 1990, p. 178). Butler rejects the idea that there is some true, biological sex onto which gender is socially added, arguing that biological understanding is socially constructed as well (Butler, 2004). Just as Goffman argues that there is no aspect of the self-untouched by the social world, Butler argues that there is no aspect of gender that can be meaningfully separated from social understanding (Lawler, 2015).

In the context of online interactions, scholars of gender have been particularly interested in the ability of online users to easily vary their gender performance (Bruckman, 1993; Curtis, 1992; Lehdonvirta et al., 2012; Morgan et al., 2020; Pagnucci & Mauriello, 1999; Ratan et al., 2019; Savicki et al., 1996; Witmer & Katzman, 1997). While this flexibility may have particular benefits for transgender users (Kitzie, 2018; Marciano, 2014; Morgan et al., 2020), the ability to temporarily escape gender norms can be beneficial to cisgender users as well (Lehdonvirta et al., 2012; Pagnucci & Mauriello, 1999).

We consider the ability of users to control when and how they perform gender as the platform affordance of "gender visibility." Platform affordances are the "possibilities for action" users see when interacting with an online interface (Evans et al., 2017; Treem & Leonardi, 2013). These affordances go beyond the technical features of a platform (i.e., the ability to create a profile) to more richly engage with the diverse and creative ways in which people actually use a platform. In other words, as people choose their avatars, complete their bios, and add their names to their online accounts, they may see an explicit opportunity to perform, change, or obfuscate perceptions of their gender. Previous work has found that users have a variety of motivations for taking these actions. Some may use online spaces as an opportunity to explore their gender identity (Kitzie, 2018; Morgan et al., 2020). Cisgender men may present as women in order to express vulnerability (Lehdonvirta et al., 2012) or to sexualize women (Curtis, 1992). Cisgender women may choose to present as men to avoid online harassment (Bruckman, 1993; Pagnucci & Mauriello, 1999). Both binary and nonbinary transgender users may similarly choose to minimize gender signals in order to evade gender-based harassment (Kaltiala & Ellonen, 2022; Lubitow et al., 2017; Waite, 2021).

Studying gendered phenomena online

All of this raises difficult questions about how researchers should study gendered phenomena in online settings. The examination of gender bias is critical, yet explicating an individual's gender is fraught. Throughout the literature, researchers have engaged a number of strategies for studying gender-based experiences, which we separate into self-report and inference methods.

Self-reports include surveys (Gosse et al., 2021; Mitchell & Štulhofer, 2021; Nadim & Fladmoe, 2021) and in-depth interviews (Chen et al., 2020; Koirala, 2020), which allow participants to self-report their gender. While a self-report is the most accurate measure of gender a researcher can get, it is important to note that these methods are also imperfect. For one thing, they reflect another kind of performance, with respondents potentially incentivized to report the gender they expect researchers are looking for. Furthermore, surveys often artificially restrict subjects to choosing between binary categories of gender. Even if additional categories are available, these may ultimately be dropped or aggregated in analysis.

Inferential methods, which involve manually or computationally estimating user gender, are generally more prevalent

Table 1. Common methods of gender inference and key conceptual dimensions differentiating these constructs

	Administrative records	Linguistic—pronouns	Linguistic—gendered terms	Linguistic—name	Handcoding
Is a subject intentionally signaling their gender categories?	Yes	Yes	Varied	No	Varied
Can a subject freely define their gender categories?	No	Yes	Yes	No	No
Can a subject easily edit their gender expressions?	No	Yes	Yes	Varied	Varied
What is gender inference primarily based on?	Recorded gender	Gender commonly associated with pronouns	Gender commonly associated with terms	Gender commonly associated with name under given cultural context	Human perception of gender performance

in the literature. The popularity of these methods comes in part from the rise of digital trace data—such as social media posts and engagement statistics—which allow researchers to directly observe online behaviors (Meyer et al., 2023; Salganik, 2017). While working with such data provides a record of who posted what and how many interactions content receives, it typically comes at the cost of having less information about the real-world users behind that behavior. In other words, researchers using digital trace data are often forced to make inferences about the people they are studying. When it comes to gender, there are several strategies researchers use to try to infer this subject characteristic. We summarize these strategies and their key conceptual dimensions in Table 1.

One method is to infer matches between users captured by digital trace data and administrative records (Barberá, 2015; Hughes et al., 2021; Pagnucci & Mauriello, 1999). For example, researchers may match voting records, school records, or other files to social media accounts. These administrative records typically include a self- or guardian report of gender, and researchers can use name or other fields to infer which records belong to which users. While powerful, such matching also comes with many limitations. Administrative records almost always restrict users to a binary selection, the recorded gender may be outdated by the time analysis takes place, and mismatches may induce error. Furthermore, matching to administrative records is time-consuming and technical, leading to infrequent use in the literature.

Perhaps the most common method of gender inference relies on automatic classification of users' linguistic signals. This includes efforts to infer gender from users' names (Liu & Ruths, 2013; Nilizadeh et al., 2021; Sebo, 2021), comments (Alipour et al., 2019; Kapidzic & Herring, 2011), or pronoun usage (Jiang et al., 2022). An additional related line of work relies on user images, sometimes in conjunction with text-based features (Messias et al., 2017; Sakaki et al., 2014). While widely used, these approaches are often deeply flawed, with name and photo inference particularly prone to racial and cultural bias in gender assignments (Pinney et al., 2023; Scheuerman et al., 2021; Sebo, 2021). Of these, we focus on pronoun usage as the most meaningful signal, as this indicates an intentional gender performance from the user.

Throughout the literature, hand-coding is often taken to be the gold standard for gender inference (Döring & Mohseni, 2019; Elias & Gurbanova, 2018; Savicki et al., 1996; Witmer & Katzman, 1997). In this approach, one or more humans make an informed judgment as to the gender of an account. This is generally more flexible and nuanced than rigid automated methods, as coders can use a range of explicit and implicit signals to inform their judgment. Hand-coding is also generally assumed to be a relatively "easy" task in which human coders can "tell" someone's gender by looking at them. While some papers rely solely on hand-coded gender inference, a common approach is to use this gold standard of hand-coding on a subset of data to justify applying automated linguistic inference to a larger dataset.

We argue that these inference methods do not capture the same constructs, and researchers must therefore be thoughtful about the specific measures relevant to their work. For example, when studying bias in how people are treated online, *perceptions* of user gender—while possibly wrong—may be the most relevant. Indeed, users who do not perform their gender in a cis-heteronormative way may be particularly likely to be misgendered—and treated as a different gender—by their online peers. These users may also be more likely to face harassment due to perceptions of their gender (Kaltiala & Ellonen, 2022; Lubitow et al., 2017; Waite, 2021). Furthermore, even binary cisgender users may leverage the affordance of gender visibility in order to change or obfuscate how their gender is perceived (Bruckman, 1993; Lehdonvirta et al., 2012; Pagnucci & Mauriello, 1999). In short, by capturing different constructs, we expect inference methods to have downstream consequences for work focused on ameliorating online gender disparities.

It is therefore essential that researchers explicitly conceptualize and name the gendered constructs they are measuring. For example, in their study of usernames and profile images, Nilizadeh et al. (2021) clearly state that they are estimating "perceived gender." Tucker and Jones (2023) are similarly careful not to conflate the use of specific pronouns with an explicit gender. Yet too often in the literature, gender inference is interpreted as a direct measure of user gender, with little nuance or gender theory applied.

Data and methods

Our analysis relies primarily on a dataset of 1.6 million Twitter users whose accounts have been matched to public U.S. voter records (Grinberg et al., 2019; Shugars et al., 2021). Data

collection and matching was approved by Northeastern University's IRB, and all data are stored on a secured server to ensure subject privacy. Using this dataset, we (1) compare how gender is expressed differently across administrative records, linguistic signals, and hand-coding of user accounts and (2) examine the downstream consequences of that variation as it relates to interpreting gendered patterns in online attention and amplification.

Matching was done in 2017 by comparing the names and locations associated with Twitter profiles to voter records compiled by the data vendor TargetSmart. Records were considered a match if a single Twitter profile matched the full name, city, and state of the individual voter record. For more details on the data collection and matching process, see Grinberg et al. (2019). For the purposes of this study, we restrict our focus to the 581,983 panel members who were active between January 1, 2021 and December 31, 2021 and had a non-blank bio and display name. Because this matching process requires users to have used their legal name on Twitter, this dataset may underrepresent transgender users, who may be less likely than average to use their legal name.

We therefore supplement portions of our analysis with a sample of accounts taken from Twitter's Decahose, a 10% sample of all tweets. Specifically, for the 11th of each month from January to December 2021, we collected all English-language tweets available through the Decahose for that day. We sampled one day per month to ensure our Decahose sample spanned the same year-long time frame as our panel data. We chose the 11th of each month as our sample day to ensure that National Coming Out Day (October 11) would be included within our data. While this could potentially result in an overrepresentation of the queer community in our Decahose sample, the scarcity of work examining genders beyond the binary made us particularly eager to ensure this population was reflected in our data. We select all users with at least one tweet in this sample and with a non-blank bio and display name and sample 10% of these accounts for computational tractability. In total, our Decahose sample includes 2,303,526 unique accounts.

Ethical considerations

Inferring gender and manually inspecting user content both introduce ethical concerns worthy of explicit consideration. First, any effort to infer user gender will result in researchers misgendering some users. This harm may be particularly acute for both binary and nonbinary transgender users who may be misgendered more frequently and experience greater levels of harassment due to their gender (Kaltiala & Ellonen, 2022; Lubitow et al., 2017; Waite, 2021). Nonbinary individuals, in particular, may self-describe and create new genders in myriad ways that are richly complex and difficult to categorize (Guyan, 2022) and may represent populations who cannot or do not want to be classified (Keyes et al., 2021). In this sense, the very act of aiming to infer users' gender may be deeply concerning (Billard, 2019; Keyes et al., 2021). However, we believe that there is value in both studying online gender disparities and in examining the downstream implications of commonly applied inference tasks. Furthermore, we believe that this analysis has particular value to the transgender community as our findings highlight the systematic erasure and misclassification these populations often face in gender-based analysis.

Second, our study involves very personal user data. While our analysis relies on publicly available social media data and voting records, matching these datasets raises ethical concerns. We take these concerns very seriously and store all data on a secured server accessible only to members of the research team. Furthermore, we only conduct, report, and share aggregate-level analysis of this matched data and will not make the underlying data publicly available. At no point did any member of the research team evaluate individual-level variation across our gender inference methods. Our hand-coding, however, required direct inspection of user content, and most people do not expect their online content to be used for research purposes (Fiesler & Proferes, 2018). However, we believe that there is significant value in our comparison of hand-coding to other gender inference methods, particularly since hand-coding is often taken to be the best method to infer gender. In order to maximize subject privacy during this study, all coding was done by members of the research team, and we restricted our coding to tweets that were publicly available at the time coding took place.

Methods for inferring gender

Administrative records

The panel approach provides our administrative measure of gender, as legal sex for most users is recorded in the voter file. Importantly, this record does not mean we have an accurate measure of each user's gender: Many states restrict citizens to a single, binary gender category, and transgender citizens may not have updated the sex associated with their voter file. Furthermore, this is an optional field in some states, and legal sex is unknown for 4.5% of our panelists. Nevertheless, administrative records are often used as measures of gender (Pagnucci & Mauriello, 1999; Shugars et al., 2021) and therefore serve as a meaningful point of comparison for our inquiry.

Linguistic signals

We focus here on the use of pronouns and gendered terms in users' bios and display names. To do so, we construct separate keyword lists for gendered terms ("mother," "father") and common pronouns ("he," "she," "they"). Both lists can be found in the Supplementary material. For all users in our data, we tokenized the text in their bio and display name and identified whether any gendered terms were present.

One challenge of working with linguistic signals is that the mere presence of a term does not necessarily indicate an intentional expression of gender. For example, someone may include the single word "they" without intending this to be a signal of gender (Tucker & Jones, 2023). Similarly, someone may indicate they are a "partner to a brilliant wife" without intending to indicate that they are a woman. We are able to overcome this challenge with respect to pronouns by following the approach of Tucker and Jones (2023) and tokenizing using the regular expression "[^a-zA-Z0-9/''-]." This method results in tokens that combine words with forward slashes, apostrophes, backticks, or hyphens, such as the pronoun series "they/them" or "they-them." While this tokenization works well for identifying a series of pronouns, it cannot be meaningfully extended to gendered terms, which are rarely expressed as a character-separated list. We therefore expect gendered terms to provide a noisier signal for gender inference than properly tokenized pronouns. Nevertheless, we see value in comparing and evaluating both these forms of linguistic signals.

Table 2. Prevalence of pronouns and gendered terms in users' bios and display names for both the panel and Decahose sample

Measure	Panel (%)	Decahose (%)
Pronouns alone		
She/her series pronouns	2.76	3.35
He/him series pronouns	1.45	1.35
They/them series pronouns	0.15	0.41
Mixed pronouns (e.g., "he/they")	0.22	0.75
Any pronoun combination	4.58	5.87
Gendered terms alone		
Only "female" terms	9.76	2.36
Only "male" terms	7.26	2.74
Both "female" and "male" terms	0.66	0.16
Any gendered term	17.70	5.29
Pronouns + gendered terms		
Female term or she/her series	12.10	5.51
Male term or he/him series	8.52	3.96
Any gendered term or pronoun	21.67	10.83
No gender indicators	78.33	89.27

Hand-coding

We hand-coded a subset of 5,064 panelist accounts to better estimate how others—both Twitter users and researchers—perceive users' gender online. In order to ensure a balance of accounts who make their gender visible in different ways, we constructed our sample for hand-coding by first splitting our panel dataset into (1) accounts whose bios included pronouns, (2) accounts whose bios included gendered terms ("mother," "father"), and (3) accounts with no linguistic gender signals (i.e., accounts not in either List 1 or List 2). Lists 1 and 2 were not mutually exclusive as some users included both pronouns and gendered terms in their bio. List 3 was by far the most common, with almost 80% of users giving no linguistic signals of gender (see Table 2). From each of these three lists, we then selected the 2,000 users who produced the most tweets in 2021. We adopted this selection strategy because active users have the largest impact on the platform and are the most frequently studied group. Since some users included both pronouns and gendered words in their bios, selecting the most active users from each of our three lists resulted in a set of 5,064 unique users whose accounts we hand-coded.

For each user in this dataset, coders were shown their most recent live tweet and asked: (1) "If you had to guess, what is the most likely gender of this user?" (Male, Female, Nonbinary, Not sure) and (2) "If you had to guess, do you think this person is transgender?" (Yes, No, Not sure). The first question gets to our core interest in perceived gender, while the second question has the potential to add meaningful context and lay the groundwork for future research on gender gaps faced by transgender individuals. To replicate the experience of scrolling through Twitter and making gendered assumptions about accounts, we used Twitter's OEmbed API to show live tweets as they would be displayed online. Coders could see the user's photo, name, and handle, along with the contents of a single tweet.

We had four coders in total, and we began with an initial subset of 200 accounts coded by all four coders. After finding that intercoder reliability across these 200 accounts was relatively high (Krippendorff's Alpha: 0.87), we split into teams, and the remaining 4,864 accounts were coded by two members of the research team. Each account was coded by one cisgender and one transgender individual. We discuss the results of our coding and intercoder reliability further in the Results section.

Measuring attention and amplification

To evaluate downstream effects of gender inference, we collect three different measures of attention and amplification for users in our panel: (1) average number of retweets per tweet, (2) average number of likes per tweet, and (3) total number of followers. Previous binary analyses lead us to expect that all these measures will be higher for accounts identified as men than for accounts identified as women (Messias et al., 2017; Nilizadeh et al., 2021; Shugars et al., 2021). While little work has focused on users beyond the gender binary, we expect these users will receive less attention than men and possibly less attention than women as well. Furthermore, we expect that variation in gender inference methods may lead to variation in downstream results and findings.

We extracted all measures from tweets collected through the Twitter API. Tweets typically gather most of their interactions within a few hours after being posted (Bae et al., 2014; Mathews et al., 2017; Pfeffer et al., 2022; Shugars & Beauchamp, 2019; Yin et al., 2021). Hence, we feel reasonably confident that our collection strategy is sufficient to ensure our data include the majority of likes and retweets these posts received. We calculate average measures of likes and retweets based on all tweets authored by a user in 2021, and we use the follower count included in that user's last tweet of 2021.

Results

Challenges of gender inference

Linguistic signals

Perhaps the most commonly used approach in the literature, linguistic signals present a computationally efficient method for inferring gender. Among the linguistic signals that might be used for this task, pronouns seem particularly promising as they reflect users' own choices to make their gender visible. Table 2 summarizes the distribution of pronouns and gendered terms within our panel and Decahose data. All pronoun measures represent combinations of the stated pronoun series. For example, the "she/her" series captures users whose bios or display names include the tokens, "she/her," "she/her/hers," "her/hers," or other relevant combinations. The full list of pronouns aggregated for each series and the gendered terms used are included in the Supplementary material.

We find that the vast majority of Twitter users (78% for the panel, 89% for the Decahose) do not use pronouns or gendered terms in their bio or display name. The prevalence of pronoun usage is relatively consistent across samples, though the Decahose, as expected, has more accounts that use either they/them or mixed (e.g., "he/they") pronouns. Only about 0.4% of our panel used these pronoun combinations, compared to 1.2% of the Decahose sample. This may be because the matching requirement for inclusion in our panel—that users associate their legal name with their Twitter account—systematically undercounts users who do not conform to a gender binary. As reported in the Supplementary material, the distribution of pronouns in both our samples is also consistent with the findings of Tucker and Jones (2023), who use Twitter's 1% sample. Notably, users in our panel appear much more likely to use gendered terms

than users in the Decahose sample (18% for the panel, 5% for the Decahose). This may also be an artifact of the panel, as our panel is restricted to those old enough to vote in 2017.

Interestingly, we find that she/her series pronouns (~3% of users) seem to be much more common than he/him series pronouns (< 1.5% of users). This prevalence gap is repeated in our Decahose sample and is consistent with the findings of Tucker and Jones (2023). This result is surprising since women typically receive more gendered harassment in online spaces and, therefore, may be less inclined to make their gender visible to others. This prevalence gap may be related to Twitter's importance for professional visibility, may reflect women refusing to obfuscate their gender despite harassment, or may indicate that cisgender women are more likely to use pronouns to signal allyship with the transgender community. Examining motivations for pronoun usage falls beyond the scope of our current study, but this is an area ripe for future research.

For the subset of accounts used in our hand-coding, we further compared the linguistic signals present in their 2021 bios to the linguistic signals present in their 2023 bios. While 91% of users signaled the same gender or had no linguistic signals in both time periods, about 9.4% of users changed the gender signaled by the words or pronouns they used. Of these, the majority (61%) are explained by users removing gendered signals from bios, while another 28% reflect users adding gendered signals after having none in the initial time window. The remaining 11% reflect users who updated the linguistic signals in their bios to indicate a different gender. While this represents an exceedingly small portion of our sample (49 users, 1% of our hand-coded sample), these numbers again reinforce that gender is not a fixed, static construct and may actively change over time for some members of the population. More research on gendered phenomena should focus on these marginalized populations in order to better understand their gendered experience online.

Hand-coding

Commonly assumed to be an easy task that is frequently held as the gold standard for gender inference, we found this measure to be flawed. While the challenges of observationally inferring gender are not surprising from the perspective of gender studies, a large portion of the computational literature in this space assumes a user's gender will be easily observable. Our findings push back against that assumption and reinforce gender as a complex construct.

Specifically, while our intercoder reliability for gender was overall quite high (Krippendorff's Alpha: 0.86), we find notable cases of disagreement. For 8% of our hand-coded sample (414 users), coders disagreed about a user's gender. For an additional 3% (133) of users, coders merely agreed that they were "not sure" of that user's gender. In total, we were unable to reliably hand-code gender for about 11% of our sample. This may represent users who choose to minimize the gendered signals associated with their account as well as users who performed their genders in ways that were not interpretable by our coders. While hand-coding may still represent a reasonable measure of gender *perception*, more work should explore the experiences of users who may be inconsistently gendered by their online peers.

Additionally, it is important to note that just because coders agree does not mean their coding is *correct*. One of the benefits of our research design is that we can compare coding not only between humans but between different inference approaches. In our hand-coded data, a mere 15 users (0.3%) were inferred to be nonbinary. Using linguistic measures, more than twice as many (31 users, 0.6%) of those same users were identified as nonbinary. Since linguistic signals capture self-expressed gender while hand-coding captures perceived gender, this may suggest that human coders are particularly bad at identifying nonbinary genders—perhaps incorrectly labeling them with a binary gender or disagreeing on which gender label to apply. Similarly, we find that human coders are extremely bad at identifying gender signals beyond a cisgender norm, as illustrated by the low intercoder reliability measure for our second question in which coders were asked to guess whether or not a user was transgender (Krippendorff's Alpha: 0.25). These findings were consistent across coding pairs and reliability measures, as reported in the Supplementary material.

We engaged in our coding as a reflexive process, aiming not only to highlight the methodological shortcomings of this approach but to illustrate the theoretical challenges. Coders universally indicated that explicitly labeling someone's gender was an inherently uncomfortable task that was more difficult than they had expected. The coding process forced the implicit to become explicit—reflecting the myriad ways gender tends to be assumed based on stereotypes and implicit signals. Coders reported relying upon several demographic and cultural signals, including name, age, interests, and "personality," to infer users' gender. For example, if an account had a traditionally masculine name and a profile picture of a sports logo, that account might be labeled as "Male" even though the coder could not see what the user looked like. Coders also reported difficulty interpreting gender through different cultural frames. While this study has not taken an intersectional approach to examining gender, it is important to remember that gender might be performed or interpreted differently across different racial or ethnic groups. Three of our coders were white and one was Asian, and our perceptions of users' gender were likely biased by our own backgrounds and frames.

Categorizing the non-categorical

A fundamental challenge to trying to infer discrete genders from online users is that gender is a complex, changing, and often non-categorical construct (Guyan, 2022; Keyes et al., 2021). We see this complexity across our measures of gender and consistently find that both binary and nonbinary transgender people are most likely to fall through the cracks of gender inference. Only about 22% of users in our dataset provide linguistic indicators of gender through pronouns or gendered terms. Of those that do, the gender inferred from linguistic signals matches the legal sex reported in the voter file 87% of the time. While this is fairly high agreement overall, it is notable that a non-trivial number of people appear to defy the gender categorization of these methods. About 26% of disagreements (4,352 users) come from people whose accounts have linguistic signals but whose gender is not recorded by the voter file. Another 35% of disagreements (5,831 users) arise from people whose linguistic signals do not correspond to genders in the voter file–for example, people using they/them pronouns or whose linguistic signals correspond to multiple genders. The remaining 38% of disagreements (6,207 users) come from people who express one binary gender through linguistic signals but have a different binary gender recorded in the voter file. If we restrict this analysis to the 5% of accounts that use pronouns, we

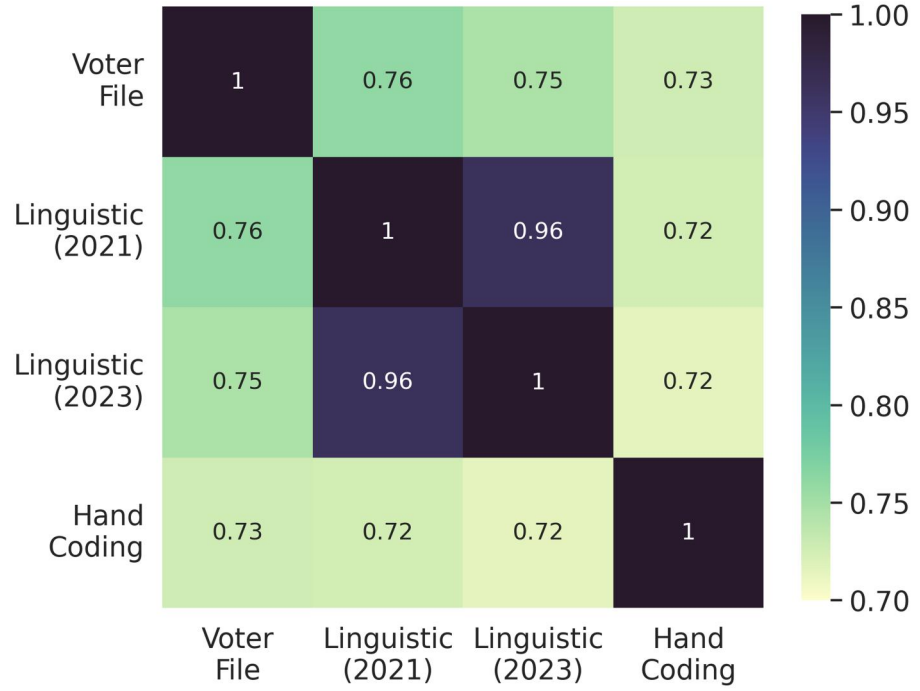

Figure 1. Pairwise agreement (Krippendorf's Alpha) between gender inferred from the voter file, linguistic signals in 2021 bios, linguistic signals in 2023 bios, and hand-coding.

similarly find that gender inferred from pronouns agrees with the voter file 86% of the time. In this case, over 57% of disagreements come from accounts indicating they/them (864 users) or mixed (1,297 users) pronouns. Another 13% of disagreements (486 users) arise from people using she/her or he/him pronouns but whose gender seems to be outdated or misrecorded in the voter file.

Figure 1 shows the pairwise agreement (Krippendorff's Alpha) between each gender inference method across our hand-coded dataset. We include estimates of gender based on linguistic signals in 2021 and 2023 bios, the binary gender indicated in the voter file, and hand-coding of 2023 profiles. While we see substantial agreement between these methods, it is notable that hand-coding—often assumed to be the gold standard—has the lowest agreement rate of any of the measures. Our hand-coding agreed with the voter file 84% of the time, with 67% of the disagreements reflecting instances where the coders themselves did not agree or were not sure of the user's gender—suggesting these users may have aimed to minimize gender signals. This disagreement is split evenly between accounts identified as male in the voter file and those identified as female, suggesting that there may not be binary differentials in this tendency. Our hand-coding also agreed 84% of the time with linguistic signals, however, the corresponding Krippendorff's Alpha is lower than for the voter file.

This may, in part, reflect our decision to code at the tweet level, as coders were instructed not to take users' bios into account. However, coding at the tweet level better replicates the user experience of scrolling through Twitter and aimed specifically to capture perceptions of gender. The relatively low performance of hand-coding demonstrates that *perceived* gender does not always align with *actual* gender. While this will be unsurprising to those familiar with theoretical understandings of gender, this disconnect is rarely explicitly acknowledged in studies using these gender inference methods.

Downstream effects in studying attention and amplification

Next, we turn to the downstream impacts gender inference methods can have. We focus specifically on examining gendered patterns in attention and amplification. Consistent with previous binary analyses (Messias et al., 2017; Nilizadeh et al., 2021; Shugars et al., 2021), we find some evidence to suggest that men typically receive more engagement with their content than women. However, the magnitude and sometimes the direction of these gaps vary depending on the gender inference method. Additionally, we find consistent evidence that users who do not fall within gender binaries never achieve the levels of attention afforded to the most popular binary accounts.

For each inference method, we report one measure of amplification (average retweets) and two measures of attention (average likes per tweet and number of followers). A summary of means and medians are reported in Table 3, and full distributions for pronoun usage and hand coding are shown in Figures 2 and 3. We estimate the statistical significance of differences in calculated distributions using the Mann–Whitney U-test (Mann & Whitney, 1947), a non-parametric test of whether random samples from one dataset are consistently higher than random samples from another dataset. This U-test is comparable to a *t*-test but is typically better when the data follow a heavy-tailed distribution (Fay & Proschan, 2010).

Administrative records

Using sex as reported in the voter file returns the result expected from previous work: Women consistently receive less attention and amplification than their male counterparts, with statistically significant differences for the average number of likes ($p < .001$) and retweets per tweet ($p < .001$). Notably, no other genders are reported in the voter file, so this finding is limited to binary analysis.

Table 3. Gender gaps in attention and amplification across different strategies for inferring users' gender

		Average retweets per tweet		Average likes per tweet		Followers		Number of tweets	Number of users
		Mean	Median	Mean	Median	Mean	Median	Median	*N*
Pronouns	She/her	0.72	0.10	7.78	2.13	1,581	395	229	16,058
	He/him	0.61	0.09	6.45	2.00	1,879	377	347	8,450
	They/them	0.65	0.07	6.32	1.95	1,413	261	303	867
	Mixed	0.58	0.08	5.96	2.04	1,094	312	340	1,297
Gendered Terms	Female	0.32	0.00	3.14	0.71	983	174	25	56,827
	Male	0.36	0.02	3.53	0.85	1,212	197	42	42,278
No pronouns or gendered terms		0.43	0.01	3.67	0.74	1,170	212	31	455,874
Voter file sex	Female	0.37	0.00	3.34	0.74	1,024	210	27	280,864
	Male	0.45	0.02	4.12	0.82	1,298	211	42	274,915
	Unknown	0.55	0.02	4.25	0.86	1,387	240	35	26,204
Hand-coded	Female	0.42	0.07	3.96	1.43	1,541	365	98	2,162
	Male	0.58	0.06	4.45	1.27	1,721	322	99	2,126
	Non-binary	0.42	0.05	4.01	1.29	664	299	913	15
	Not sure	1.06	0.06	11.20	1.30	1,528	238	99	123
	Mixed	0.45	0.05	4.93	1.28	907	233	99	396

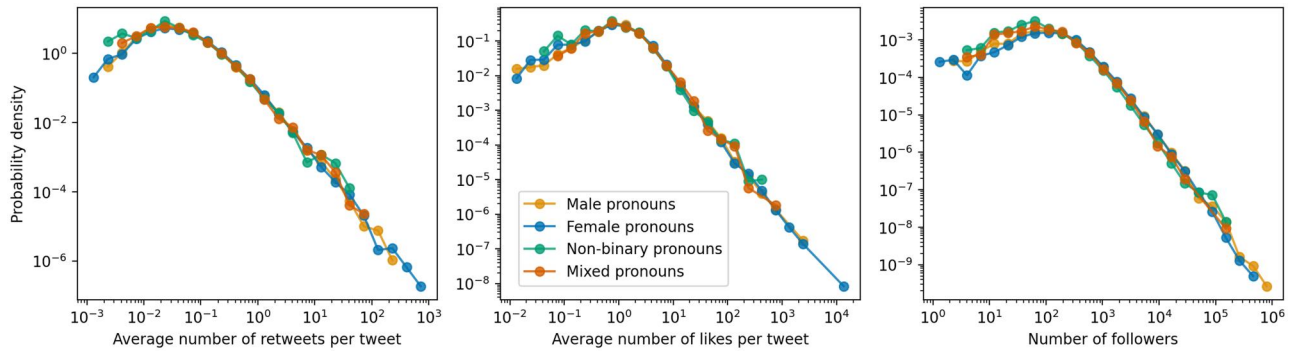

Figure 2. Distribution of attention and amplification using linguistic features. Log-log scale.

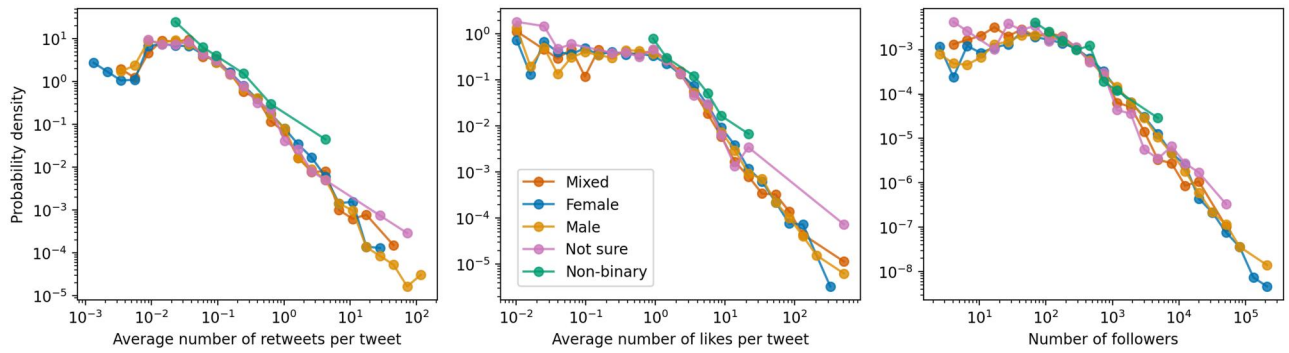

Figure 3. Distribution of attention and amplification measures using hand coding inference. Log-log scale.

Linguistic signals

Interpretations of this gender gap are sensitive to the type of linguistic signal used: Accounts that use "female" gendered terms receive lower attention than those who use "male" gendered terms ($p < .001$ for all three measures). In contrast, people who use "she/her" pronouns actually tend to receive *more* attention than those who use "he/him" pronouns ($p < .001$ for likes and retweets, $p < .01$ for followers). Additionally, we find that users with "they/them" or mixed pronoun combinations receive significantly less attention than users with "he/him" or "she/her" pronouns ($p < .01$ for all three measures), suggesting that nonbinary populations may face unique marginalization patterns which are not captured by binary analyses.

Throughout this analysis, it is important to remember that only 20% of accounts include linguistic signals, raising interesting questions about what motivates users to perform their gender in this way. While the question of motivation falls beyond the scope of our current study, we can observe some systematic ways in which those who use linguistic signals are not typical of average users. Those who use pronouns tend to be more active and receive more attention than other users

($p < 0.001$ for all three measures). This population may be different from average Twitter users in other ways as well, with previous work showing that those who use pronouns are substantially more likely to follow each other and share similar political interests (Jiang et al., 2022; Tucker & Jones, 2023). We also find a notable age differential between those who use pronouns and those who use gendered terms: the median age of panelists who include pronouns is 35, while the median age of those with gendered terms is 46. All of this points to rich areas of inquiry that open when a more nuanced understanding of gender is considered.

Hand-coding

With this inference method, we again see complicated gender dynamics around attention and amplification. Users coded as "female" appear to get more likes and have more followers than those coded as "male" ($p < .05$ for likes and followers, $p = .06$ for retweets). Furthermore, we see that this trend reverses if we use the mean rather than the median for this calculation. In other words, while users coded as "female" tend to receive more engagement than those coded as "male," the most popular "male" users received sizably more attention than their "female" counterparts. While we are unable to make strong claims about the experiences of nonbinary users ($n = 15$), our findings suggest that this population, as well as users where coders were "not sure" of the gender ($n = 123$) or did not agree ($n = 391$) typically received less attention and amplification than users coded as "male" or "female" (retweets, $p < .05$, followers, $p < .001$). These findings again suggest that users who are not perceived to be part of the gender binary may tend to get less attention than their binary counterparts.

Discussion

There is tremendous value in studying gendered phenomena in online spaces, and a rich line of research has aimed to understand and ameliorate gender gaps in who receives attention and whose voice is amplified (Gallagher et al., 2021; Jackson et al., 2020; Nilizadeh et al., 2021; Shugars et al., 2021). Yet, this work necessarily requires categorizing the non-categorical—in order to examine and compare gendered experiences and phenomena, researchers must infer or otherwise estimate user gender.

While researchers may choose from a number of popular strategies for this task, no method can overcome the fundamental theoretical challenge facing this work: Gender is a complex social construct that may be performed differently in different contexts (Butler, 1990, 2004; Guyan, 2022; Keyes et al., 2021). In online settings, the affordance of gender visibility introduces particular challenges, as users can easily choose how to perform, change, or obfuscate their gender (Kitzie, 2018; Lehdonvirta et al., 2012; Marciano, 2014; Morgan et al., 2020; Pagnucci & Mauriello, 1999; Ratan et al., 2019).

The practical result of these challenges is that most studies of online gender phenomena adopt a binary, cisgender approach: identifying users who are perceived to be "male" or perceived to be "female" and either dropping or aggregating accounts that do not perform gender in line with these norms. Nonbinary users are definitionally erased in these studies, while both cisgender and transgender binary users may similarly be ignored or misclassified if their gender performance does not align with an expected cisgender norm.

In this article, we demonstrate the empirical implications of these theoretical challenges. Using a dataset of 1.6 million Twitter users matched to U.S. voting records, we apply three common methods for gender inference: administrative records, linguistic signals, and hand-coding. We demonstrate that how gender is inferred can have significant downstream consequences on the interpretation of gendered phenomena online, showing specifically that these measures suggest different gendered patterns in attention and amplification. Our empirical results consistently reinforce theoretical understandings—gender is a complex construct that may change over time and be expressed in different ways by different populations. While binary classification may accurately reflect a majority of the population, people not represented by those categories are consistently under-studied and may face some of the largest disparities in gendered treatment.

Consistent with past research (Messias et al., 2017; Nilizadeh et al., 2021; Shugars et al., 2021), we find evidence to suggest that women tend to receive less attention than their male counterparts, though this finding is primarily demonstrated through the use of administrative records to infer gender. Measures relying on linguistic signals and hand-coding found more complex and, at times, contradictory results. For example, people who indicate "she/her" pronouns tend to receive more attention than those who use "he/him" pronouns. While pronoun detection presents a particularly promising method of gender inference, as it reflects user self-expression and is computationally tractable, these findings suggest that users who include pronouns in their bios may be systematically different from those who do not. More broadly, our findings suggest that performances of gender may not only affect gender inference but may have downstream consequences for interpreting gendered phenomena online. Any research relying on inferred gender should therefore explicitly articulate the construct they believe they are measuring and assess how the inferential methods selected may affect study outcomes.

Perhaps most importantly, our findings consistently suggest that all these methods risk undercounting nonbinary users and others who do not perform gender in accordance with a cisgender norm. Nonbinary users, in particular, are not recorded in the voter file and appeared to be undercounted in our hand-coding. While linguistic signals, particularly pronouns, have the benefit of reflecting a self-reported measure, less than a third of users employ pronouns or gendered terms in their bios, making this an unfeasible standalone measure for describing the larger population. In measures where nonbinary genders are recorded, this population appears to receive consistently less attention than binary users, suggesting a pressing need for further research on gender disparities facing these users.

We additionally find that hand-coding is unlikely to be a panacea for gender inference: How a user's gender is perceived does not necessarily reflect that person's actual gender. This suggests that researchers should think carefully about what constructions of gender are most likely to shape the gendered phenomena they study. If researchers are primarily interested in how *perceptions* of gender drive variation in user treatment and experience, hand-coding of those perceptions may be appropriate. In these cases, researchers should be clear that their method aims to infer perceptions, which may not match actual user gender. If a research question

requires estimating users' actual gender, administrative records supplemented by linguistic signals are probably the best methods. When evaluating the matching between our three methods, we find that our administrative measure and linguistic signals are in better agreement than hand-coded gender. This is notable since hand-coding captures perceptions of gender, while the other two measures reflect some form of self-report.

There is no accurate way to infer gender or to perfectly categorize this non-categorical construct. Yet, the work of investigating bias, disparities, and other gendered phenomena online is critical. In this article, we do not aim to "solve" the unsolvable problem of gender inference. Rather, we have aimed to highlight the empirical implications of gender theory by demonstrating the downstream consequence of common gender inference techniques. Our contribution is in problematizing these measures. Our findings suggest that researchers need to think critically about how they operationalize gender and interpret results based on these measures. Furthermore, our findings consistently suggest the need for more research explicitly focused on the gendered experiences of nonbinary users and those who perform gender outside a cisgender norm. In short: More work must be done to better analyze and address gaps beyond the gender binary.

Supplementary material

Supplementary material is available online at *Journal of Computer-Mediated Communication*.

Data availability

The data underlying this article cannot be shared due to privacy concerns arising from matching data to administrative records. Secondary data and aggregated analysis are publicly available at http://sarahshugars.com/gender-categorization.

Conflicts of interest: None declared.

Funding

Funding support for this article was provided by the Volkswagen Foundation.

Acknowledgements

The authors would like to thank the special issue editors, Sandra González-Bailón and Ágnes Horvát, as well as the anonymous reviewers for their helpful and thoughtful feedback. The authors would also like to thank Casey Randazzo, who helped with some of the hand coding of accounts.

References

Alipour, B., Imine, A., & Rusinowitch, M. (2019). Gender inference for Facebook picture owners. In S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Trust, privacy and security in digital business* (Vol. *11711*, pp. 145–160). Springer International Publishing. https://doi.org/10.1007/978-3-030-27813-7_10

Armentor-Cota, J. (2011). Multiple perspectives on the influence of gender in online interactions: Role of gender in online interactions. *Sociology Compass*, *5*(1), 23–36. https://doi.org/10.1111/j.1751-9020.2010.00341.x

Bae, Y., Ryu, P.-M., & Kim, H. (2014). Predicting the lifespan and retweet times of tweets based on multiple feature analysis. *ETRI Journal*, *36*(3), 418–428. https://doi.org/10.4218/etrij.14.0113.0657

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Billard, T. J. (2019). "Passing" and the politics of deception: Transgender bodies, cisgender aesthetics, and the policing of inconspicuous marginal identities. In T. Docan-Morgan (Ed.), *The Palgrave handbook of deceptive communication* (pp. 463–477). Springer International Publishing. https://doi.org/10.1007/978-3-319-96334-1_24

Bruckman, A. S. (1993). Gender swapping on the Internet. In: *Proceedings of INET '93, San Fransisco, CA*.

Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. Routledge. https://doi.org/10.4324/9780203902752

Butler, J. (2004). *Undoing gender*. Routledge. https://doi.org/10.4324/9780203499627

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, *21*(7), 877–895. https://doi.org/10.1177/1464884918768500

Curtis, P. (1992). MUDding: Social phenomena in text-based virtual realities. *Intertek*, *3*(3), 26–48.

Döring, N., & Mohseni, M. R. (2019). Fail videos and related video comments on YouTube: A case of sexualization of women and gendered hate speech? *Communication Research Reports*, *36*(3), 254–264. https://doi.org/10.1080/08824096.2019.1634533

Elias, S., & Gurbanova, N. (2018). Relocating gender stereotypes online: Critical analysis of sexist hate speech in selected social media. *Proceedings of the International Conference on Language Phenomena in Multimodal Communication (KLUA 2018)*. International Conference on Language Phenomena in Multimodal Communication (KLUA 2018), Surabaya, Indonesia. https://doi.org/10.2991/klua-18.2018.40

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. https://doi.org/10.1111/jcc4.12180

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*(1), 1–39. https://doi.org/10.1214/09-SS051

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1), 1–14. https://doi.org/10.1177/2056305118763366

Gallagher, R. J., Doroshenko, L., Shugars, S., Lazer, D., & Foucault Welles, B. (2021). Sustained online amplification of COVID-19 elites in the United States. *Social Media + Society*, *7*(2), 1–16. https://doi.org/10.1177/20563051211024957

Goffman, E. (1955). On face-work. *Psychiatry*, *18*(3), 213–231. https://doi.org/10.1080/00332747.1955.11023008

Goffman, E. (1959). *The presentation of self in everyday life*. Anchor.

Gosse, C., Veletsianos, G., Hodson, J., Houlden, S., Dousay, T. A., Lowenthal, P. R., & Hall, N. (2021). The hidden costs of connectivity: Nature and effects of scholars' online harassment. *Learning, Media and Technology*, *46*(3), 264–280. https://doi.org/10.1080/17439884.2021.1878218

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guyan, K. (2022). *Queer data: Using gender, sex and sexuality data for action*. Bloomsbury Academic.

Hughes, A. G., McCabe, S. D., Hobbs, W. R., Remy, E., Shah, S., & Lazer, D. M. J. (2021). Using administrative records and survey

data to construct samples of tweeters and tweets. *Public Opinion Quarterly*, *85*(S1), 323–346. https://doi.org/10.1093/poq/nfab020

Jackson, S. J., Bailey, M., & Foucault Welles, B. (2020). *#HashtagActivism: Networks of race and gender justice*. The MIT Press. https://doi.org/10.7551/mitpress/10858.001.0001

Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H.-C. H., & Ferrara, E. (2022). *What are your pronouns? Examining gender pronoun usage on twitter*. arXiv 2207.10894. http://arxiv.org/abs/2207.10894

Kaltiala, R., & Ellonen, N. (2022). Transgender identity and experiences of sexual harassment in adolescence. *Child Abuse Review*, *31* (4), 1-14. https://doi.org/10.1002/car.2748

Kapidzic, S., & Herring, S. C. (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication*, *17*(1), 39–59. https://doi.org/10.1111/j.1083-6101.2011.01561.x

Keyes, O., May, C., & Carrell, A. (2021). You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23. https://doi.org/10.1145/3449113

Kitzie, V. (2018). "I pretended to be a boy on the Internet": Navigating affordances and constraints of social networking sites and search engines for LGBTQ+ identity work. *First Monday, 23(7), 1–10*. https://doi.org/10.5210/fm.v23i7.9264

Koirala, S. (2020). Female journalists' experience of online harassment: A case study of Nepal. *Media and Communication*, *8*(1), 47–56. https://doi.org/10.17645/mac.v8i1.2541

Lawler, S. (2015). *Identity: Sociological perspectives*. John Wiley & Sons.

Lehdonvirta, M., Nagashima, Y., Lehdonvirta, V., & Baba, A. (2012). The stoic male: How avatar gender affects help-seeking behavior in an online game. *Games and Culture*, *7*(1), 29–47. https://doi.org/10.1177/1555412012440307

Liu, W., & Ruths, D. (2013). What's in a name? Using first names as features for gender inference in Twitter. In: *AAAI Spring Symposium Series*, *Palo Alto, CA*.

Lubitow, A., Carathers, J., Kelly, M., & Abelson, M. (2017). Transmobilities: Mobility, harassment, and violence experienced by transgender and gender nonconforming public transit riders in Portland, Oregon. *Gender, Place & Culture*, *24*(10), 1398–1418. https://doi.org/10.1080/0966369X.2017.1382451

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Marciano, A. (2014). Living the VirtuReal: Negotiating transgender identity in cyberspace. *Journal of Computer-Mediated Communication*, *19* (4), 824–838. https://doi.org/10.1111/jcc4.12081

Mathews, P., Mitchell, L., Nguyen, G., & Bean, N. (2017). The nature and origin of heavy tails in retweet activity. In: *Proceedings of the 26th International Conference on World Wide Web Companion—WWW '17 Companion, Perth, Australia* (pp. 1493–1498). https://doi.org/10.1145/3041021.3053903

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, *47(A)*, 46–55. https://doi.org/10.1016/j.wsif.2014.07.012

Messias, J., Vikatos, P., & Benevenuto, F. (2017). White, man, and highly followed: Gender and race inequalities in Twitter. *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany* (pp. 266–274). https://doi.org/10.1145/3106426.3106472

Meyer, M. N., Basl, J., Choffnes, D., Wilson, C., & Lazer, D. M. J. (2023). Enhancing the ethics of user-sourced online data collection and sharing. *Nature Computational Science*, *3*(8), 660–664. https://doi.org/10.1038/s43588-023-00490-7

Mitchell, K., & Štulhofer, A. (2021). Online sexual harassment and negative mood in Croatian female adolescents. *European Child & Adolescent Psychiatry*, *30*(2), 225–231. https://doi.org/10.1007/s00787-020-01506-7

Morgan, H., O'Donovan, A., Almeida, R., Lin, A., & Perry, Y. (2020). The Role of the Avatar in Gaming for Trans and Gender Diverse Young People. *International Journal of Environmental Research and Public Health*, *17*(22), Article 22. https://doi.org/10.3390/ijerph17228617

Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review*, *39*(2), 245–258. https://doi.org/10.1177/0894439319865518

Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.-Y., Kapadia, A., & Rojas, F. (2021). Twitter's Glass ceiling: The effect of perceived gender on online visibility. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 289–298. https://doi.org/10.1609/icwsm.v10i1.14711

Pagnucci, G. S., & Mauriello, N. (1999). The masquerade: Gender, identity, and writing for the web. *Computers and Composition*, *16* (1), 141–151. https://doi.org/10.1016/S8755-4615(99)80010-3

Pfeffer, J., Matter, D., & Sargsyan, A. (2022). The half-life of a tweet. Preprint. http://pfeffer.at/papers/half-life.pdf

Pinney, C., Raj, A., Hanna, A., & Ekstrand, M. D. (2023). Much Ado about gender: Current practices and future recommendations for appropriate gender-aware information access. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, New York, NY* (pp. 269–279). https://doi.org/10.1145/3576840.3578316

Ratan, R. A., Fordham, J. A., Leith, A. P., & Williams, D. (2019). Women keep it real: Avatar gender choice in league of legends. *Cyberpsychology, Behavior, and Social Networking*, *22*(4), 254–257. https://doi.org/10.1089/cyber.2018.0302

Sakaki, S., Miura, Y., Ma, X., Hattori, K., & Ohkuma, T. (2014). Twitter user gender inference using combined analysis of text and image processing. *Proceedings of the Third Workshop on Vision and Language, Dublin, Ireland* (pp. 54–61). https://doi.org/10.3115/v1/W14-5408

Salganik, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press. https://academic.oup.com/jrsssa/article/181/3/917/7072048

Savicki, V., Lingenfelter, D., & Kelley, M. (1996). Gender language style and group composition in internet discussion groups. *Journal of Computer-Mediated Communication*, *2*(3), 1–10. https://doi.org/10.1111/j.1083-6101.1996.tb00191.x

Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society*, *8*(2), 1–15. https://doi.org/10.1177/20539517211053712

Sebo, P. (2021). Performance of gender detection tools: A comparative study of name-to-gender inference services. *Journal of the Medical Library Association, 109*(3), 414–421. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8485937/

Shugars, S., & Beauchamp, N. (2019). Why keep arguing? Predicting engagement in political conversations online. *SAGE Open*, *9*(1), 1–13. https://doi.org/10.1177/2158244019828850

Shugars, S., Gitomer, A., McCabe, S., Gallagher, R. J., Joseph, K., Grinberg, N., Doroshenko, L., Foucault Welles, B., & Lazer, D. (2021). Pandemics, protests, and publics: Demographic activity and engagement on Twitter in 2020. *Journal of Quantitative Description: Digital Media*, *1*(2021), 1–68. https://doi.org/10.51685/jqd.2021.002

Treem, J. W., & Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, *36*(1), 143–189. https://doi.org/10.1080/23808985.2013.11679130

Tucker, L., & Jones, J. (2023). Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among US Twitter users 2015-2022. *Journal*

of Quantitative Description: Digital Media, *3*(2023), 1–35. https://doi.org/10.51685/jqd.2023.003

Waite, S. (2021). Should I stay or should I go? Employment discrimination and workplace harassment against transgender and other minority employees in Canada's federal public service. *Journal of Homosexuality*, *68*(11), 1833–1859. https://doi.org/10.1080/00918369.2020.1712140

Witmer, D. F., & Katzman, S. L. (1997). On-line smiles: Does gender make a difference in the use of graphic accents? *Journal of Computer-Mediated Communication*, *2*(4), 1–10. https://doi.org/10.1111/j.1083-6101.1997.tb00192.x

Yin, H., Yang, S., Song, X., Liu, W., & Li, J. (2021). Deep fusion of multimodal features for social media retweet time prediction. *World Wide Web*, *24*(4), 1027–1044. https://doi.org/10.1007/s11280-020-00850-7