Much Ado About Gender

Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access

Christine Pinney
Amifa Raj
christinepinney@u.boisestate.edu
amifaraj@u.boisestate.edu
People & Information Research Team
Boise State University
Boise, Idaho, USA

Alex Hanna
alex@dair-institute.org
DAIR Institute
USA

Michael D. Ekstrand
ekstrand@acm.org
People & Information Research Team
Boise State University
Boise, Idaho, USA

ABSTRACT

Information access research (and development) sometimes makes use of gender, whether to report on the demographics of participants in a user study, as inputs to personalized results or recommendations, or to make systems gender-fair, amongst other purposes. This work makes a variety of assumptions about gender, however, that are not necessarily aligned with current understandings of what gender is, how it should be encoded, and how a gender variable should be ethically used. In this work, we present a systematic review of papers on information retrieval and recommender systems that mention gender in order to document how gender is currently being used in this field. We find that most papers mentioning gender do not use an explicit gender variable, but most of those that do either focus on contextualizing results of model performance, personalizing a system based on assumptions of user gender, or auditing a model's behavior for fairness or other privacy-related issues. Moreover, most of the papers we review rely on a binary notion of gender, even if they acknowledge that gender cannot be split into two categories. We connect these findings with scholarship on gender theory and recent work on gender in human-computer interaction and natural language processing. We conclude by making recommendations for ethical and well-grounded use of gender in building and researching information access systems.

CCS CONCEPTS

• **Social and professional topics** → **Gender**; • **Information systems** → **Information retrieval**.

KEYWORDS

information access, gender, auditing, systematic review

ACM Reference Format:
Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576840.3578316

CHIIR '23, March 19–23, 2023, Austin, TX, USA

1 INTRODUCTION

Research and development of *information access systems* (IAS) — search engines, recommender systems, and similar systems that facilitate access to information, often studied in conferences on information retrieval (IR) and related topics such as recommendation and user modeling — often engage with gender in some way or another. These uses vary, from reporting the demographic distribution of participants in a user study to using gender as a feature in personalized results to seeking to ensure the system treats users or content providers of various genders fairly, among other objectives. There has been little explicit consideration in this literature, however, about how gender should be used in information access. Most work takes gender as a categorical feature that can be obtained from users or inferred from the underlying data set and uses it as any other feature in the system. There are several important questions about the use of gender in information access research, including:

- When should gender be used, and when is it inappropriate, unhelpful, or harmful to use gender in research or practice?
- When it is appropriate to use gender, how should gender be defined and operationalized?
- Where and how should gender data be obtained? Are there methods that are best avoided?

Our goal in this paper is to document the current state of research practice with respect to these questions and provide a foundation for discussion, further research, and well-grounded practice among information access researchers, practitioners, affected parties, and others that moves the community towards thoughtful, principled use and non-use of gender. We agree that it is indeed crucial for search engines, recommender systems (RS), and other information access systems to provide effective, appropriate, and useful results to users of all genders and other demographic affiliations. We argue that this is best done through careful attention to the meaning of gender and how its use and operationalization affects the people the system is aiming to assist, particularly people with marginalized gender identities and adverse experiences with computational and datafied representations of gender.

To that end, we organize this paper in two parts. First, we provide a systematic review and analysis of the use of gender in recent publications in key information access research venues. We then identify goals for which gender is used, ways it is encoded, and the data sources used to obtain gender information for users, content providers, and other affected people. Finally, we build on this survey

and relevant literature from other domains to provide recommendations for improving research and implementation practices around gender in information access.

We are certainly not the first to question how gender is used in computing systems. Hamidi et al. [30] and Scheuerman et al. [66] have done crucial work on the (mis)use of gender in human-computer interaction, and [14] have looked at how it is used in natural language processing (NLP) research. This highlights how this issue is not unique to IAS; indeed, this is a common issue in quantitative social sciences writ large [75]. We complement their work by specifically investigating information access applications, including search and recommendation.

2 MOTIVATING VIGNETTES

The use of gender as a variable in information access systems may be becoming more ubiquitous. Gender may be used as an input to a recommender system or information retrieval model. Some of the uses of gender may present themselves as more insidious than others. To motivate our interest in understanding the use of gender, we present two vignettes.

In China, Kentucky Fried Chicken partnered with Baidu to offer a product which provided food recommendations based on details inferred from a customer's face at 300 stores in Beijing [23]. In addition to inferring gender, the facial analysis product also inferred age and "beauty" [34]. The tool recommends different meals which are seemingly based on these factors. For instance, the author of the Guardian article was read as a woman in her 30s, and the system recommended a chicken hamburger meal. A press release from Baidu suggested that "'a male customer in his early 20s' would be offered 'a set meal of crispy chicken hamburger, roasted chicken wings and coke', while 'a female customer in her 50s' would get a recommendation of 'porridge and soybean milk for breakfast'."

Gender itself is inferred in this system from gender expression, which has been criticized in the literature which we discuss below. Moreover, strong assumptions are made about the role gender should play in product recommendation. It's not clear how, *prima facie*, how these meals correlate with these inferred features. In what way does it make sense for features such as inferred gender, beauty, or age to serve as a suggestion for meal items? Are those features indicative of purchasing behavior or desired products? To us, these features, inferred from personal appearance, make spurious product recommendations. However, what we do know is that the system presents a new avenue for massive collection of facial images and purchasing patterns, which could be used by Baidu to monetize other aspects of social and economic life in China.

Another, more positive, use of gender can be found in an audit conducted by Spotify to assess how female artists are represented and made visible to listeners through the platform's discovery tools [21]. The authors of this study found that recommendations had a slightly higher proportion of female artists than users' "organic" behavior (i.e. behavior which was not recommendation-driven), and further, that recommending more female artists correlated with increases in later user-initiated streaming of music by female artists.

In this case, gender as a variable is used as an identifying feature of a recommended product. Such work can be valuable in understanding how information access technologies interact with societal discrimination, when they propagate such biases, and how they may be deployed as interventions to promote more equitable information economies [17].

3 BACKGROUND

Gender has been discussed in various ways in information access research throughout the history of the relevant fields, and there is also a rich literature on the construct of gender and its interaction with data and computation. To set the stage for our formal review in Section 4, we first briefly outline some of that background here.

3.1 The Uses of Demographics in IAS

As noted in the introduction and explored much more thoroughly in our systematic review, there are a variety of ways that gender appears in information access research. One of the earliest recommender systems, Grundy [57], explicitly used a user's gender as a component of its model of their preferences and incorporated gender stereotypes into its initial recommendations (which the user could refine through subsequent conversational interaction); in modern personalization, gender is one of the many attributes data brokers routinely collect and sell to companies to use for a variety of purposes [13]. Early work on matrix factorization for collaborative filtering used a gender affinity axis ("geared toward females" vs. "males") to illustrate the idea of embedding movies [38]. A more recent line of work seeks to understand information access systems' differential impacts to see if they are treating people of different genders "fairly" as users [20, 44], as producers of the information being retrieved [18, 21, 25], or as the subjects of that information [19, 35, 45].

Aside from discussions in limitations sections of some of these papers, there is little work on when, why, and how gender is and should be used in information access research, or putting this work in the context of discussions about gender in social science or other computing fields such as human-computer interaction. This is the gap we seek to fill in this paper.

3.2 Gender as a Category

Much of the literature within sociology and gender studies has focused on the differences between gender and sex. Typically, "sex" is used to refer to biological characteristics while "gender" is related to internal perceptions of self and how external society sees individuals. However, gender and sex are entangled, and sex itself is socially constructed by scientists, policymakers, and technologists [24, 63].

Gender scholars, as well as transgender and queer activists, have also made the distinction between gender identity and gender expression. Gender identity typically refers to one's own internal understanding of gender and self-identification. Gender expression refers to how one presents one's own gender and wants to be seen by the world. These both can fit into binary notions of gender, but can also be expansive and encompass a constellation of different identifications and notions of what self-expression can entail. Moreover, gender expression can be broken up both internally (how one is expressing one's gender and feels about it to themselves) and how others perceive that individual's gender (perceived gender expression). In this article, we follow [64] and focus on discussing

gender, given that technological artifacts and systems typically discuss social constructions of gender as datafied by informational systems. However, it is important to note that many types of information access systems may make claims about having data on sex (e.g. through medical imaging or genomics).

Gender data can be obtained in a plethora of ways, depending on the modality. There is robust literature within social science research on how to survey for gender, especially when that measurement moves beyond the male/female gender binary. In survey research, much of the focus has centered around ensuring that population-level estimates can be inferred from a sample that is attentive to the individuals who do not fit into either the category of male or female. The Williams Institute has developed tools which ask respondents if they identify within the binary and then asks about transgender status [72]. This has been criticized as being too reductive, however, and may not be applicable for smaller scale studies. Others have focused on attempting to obtain a measure of how others may perceive their gender expression [41]. More recently, many others have addressed how to approach gender as a matter of data justice using intersectional feminist and queer theory lenses [16, 29].

Gender data come from many different places in the papers we examine, so we do not distinguish between gender identity and gender expression. However, it is important to note that these two categories are used nearly interchangeably in the computer science literature that we surveyed.

3.3 Gender in Computational Research

With the rise in attention to facial recognition as a technology, many researchers within HCI and AI have focused on the attribution of gender to individual data traces, typically images of people. Keyes [36] has written on the dangers of automatic gender recognition (AGR), Scheuerman et al. [65] have written on how AGR systems perform worse on trans and gender non-conforming people, and how these systems cannot legibly recognize non-binary people. Gender non-conforming and transgender individuals also feel as though these systems produce harm by involuntarily gendering them [30]. "Gender" is also necessarily raced; that is, binary genders themselves are the endpoint of processes of centuries of European colonization [64] and erase other genders which were part of indigenous and non-Western societies. Moreover, gender assessments are typically not accorded the same status to non-white women, especially Black women, as evidenced by [8].

Moreover, although there is less academic research in the interaction of gender and text, this is still a strain of research which manifests in a few different registers. There is a body of work which attempts to predict gender from textual prose (e.g. [51], however much of the work in natural language processing focuses on the notion of gender bias in text and text representations. One of the most major of these interventions [6, 9] suggests that pre-trained embedding spaces exhibit sexist biases (e.g. doctors are to males, whereas nurses are to females). Recent work has suggested that, although there is significant work in gender bias in NLP, few of these papers engage with gender theory, consider non-binary genders, or consider the intersectional, already-racialized notion of gender [14]. At the intersection of computer vision and natural language processing, gender and racialized-gender bias persist in multi-modal domains, such as image search [49], text-image benchmarks [15], and multi-modal models such as SCAN and CLIP [74].

4 REVIEW OF CURRENT PRACTICES

In order to better understand the landscape of the use of gender in information access systems, we conducted a survey of all papers which mentioned sex or gender in key information retrieval and recommendation systems publication venues. We desired to assess what, in particular, this academic community was doing with the concept of gender in academic outputs.

4.1 Methods

To collect a set of papers to analyze, we searched for all papers that mentioned gender-related words in SIGIR, CHIIR, RecSys, UMAP, and TOIS papers in 2017-2021 using the ACM Digital Library search. We selected these venues to furnish examples representative of multiple perspectives in, particularly from a computer science perspective; papers in these venues are influential across both research and practice. We selected these years because we wanted to take a snapshot of relatively current research within this field rather than attempt to make any larger claims about changes over time, particularly extending to earlier days of information access research.

We constructed a codebook based on a sample of articles matching our criteria. The codebook was constructed at the guidance of the third author, a sociologist who focuses on the intersections of technology, race, and gender, and the final author, a senior computer scientist focusing on information access systems. New questions were added as needed. For instance, we began the study focusing online on whether there was a gender variable and the goals of using gender, with the assumption that most articles would address user gender. However, we then came to understand that gender may have different referents (e.g. the data instance), and that there may be multiple referents. Moreover, we began to find that many of the uses of gender were part of an audit process to detect bias, so we added another question regarding those explicitly.

The lead author then coded for each of the variables in our codebook across all the articles. All authors met weekly to discuss the coding process and resolve ambiguities, and to work through exemplar cases with the lead author. Because a single person coded all the articles, we do not report interrater reliability metrics. The full coding process, the codebook, and the dataset are available in citation [52].

4.1.1 Variables. For each paper, we coded for several different variables. Table 1 provides a summary at a glance.

What is the primary referent? The referent is the group of people who gender is being attributed to. This may be users of a recommender system, subjects of particular data instances (such as clothing or musical artists), or annotators who are labeling data. In cases in which the paper conducted a user study using a crowdworking platform, we coded study participants as "users." While we began this study anticipating only "users", "subjects", and "providers" being our referent, we added annotators as we continued to code.

Variable	Possible values
Primary referent	User/Subject/Provider/Annotator
Gender variable?	Yes/No
Gender categories	Binary/Binary+Other
Gender determination	Self-identification/Annotator/Inferred
Bias/fairness?	Yes/No
Goals	User study/survey
	Gender personalization
	Audit system behavior
	Gender prediction
	Protect gender variable
	Persona generation

Table 1: Summary of variables

Is there a gender variable? We determined if there was any kind of gender variable in the article text at all. If there is no gender variable, then this disqualifies answering other questions about the paper. We coded a paper as "applied" when the model or experiment in the paper did not use gender, but the authors suggest that gender could be used with their method.

What are the gender categories? This variable outlined which values the gender variable will take. Sometimes these were explicitly mentioned, but often they will be obscured in a table or implicit in a statistical model. Moreover, we also noted if the authors coded for a third gender, such as "other" or "non-binary." We also coded for whether the authors verbally acknowledged that gender was non-binary, but did not operationalize this in any way. We did so because we hypothesized that some authors would make a textual note that gender was non-binary, but then continue using binary values for gender.

How is gender determined? We coded how the authors are obtaining the gender label. The gender label itself may come from self-identification by the user, or from an inference being made by the authors, third-party annotators (such as crowdworkers), or an automated system. We began from two expected categories (self-identification and machine-inferred) but added crowdworker inference as we noticed this in the data.

Is this paper about bias and/or fairness? Many papers will be about assessing the bias with a particular system or dataset, attempting to debias a dataset, or create a fair dataset or method. This would be more akin to the auditing example noted above.

Goals of using gender. Lastly, we coded for the "goal" of the use of the gender variable. Instead of defining a set of discrete goals which gender was used for, this was an inductive category, in which we added different goals progressively. This included some goals which we expected at the start of the research project, such as "Personalize based on gender" or "Gender prediction" (both used in the KFC China example above), but also encompassed some surprising uses. We discuss these inductively coded goals below in the findings.

4.2 Overview of Data and Univariate Findings

We collected 801 papers from 4 conferences (CHIIR, SIGIR, RecSys, and UMAP) and one journal (TOIS); of these, we coded 598 papers and excluded 203 workshop summary papers that didn't have sufficient peer review to code. Of the 598 coded papers, we found that 73 papers had a gender variable of interest, 442 did not have a gender variable, and 57 had a gender variable that was "applied."

4.2.1 Gender Referent. In each paper, the authors attribute gender to a specific object — the person or thing that the authors are referring to when discussing gender. If authors attributed gender to multiple entities, one entity was labeled as the primary referent and the paper was coded as having multiple referents. We identified 4 types of referents with which authors associated gender.

User Referent (52 papers). This set of papers considers gender association of users who interact with systems [12, 33, 43, 55, 71]. This user interaction may be direct where gender identity is self-declared (user study or survey), or it may be indirect where gender identity is annotated or inferred (annotation of user-generated profile, facial inference). For example, Rozen et al. [59] used user-stated gender information to evaluate their proposed system in predicting user demographic attributes, namely gender, from user browsing data and generated comments on news articles.

Subject Referent (15 papers). In this group of papers, gender is associated with subjects or items. Gender of items can be inferred from item content, for example, song lyrics, documents, and dataset labels [4, 48, 69, 77]. For instance, Rekabsaz and Schedl [56] use gendered keywords to identify female/male magnitude of retrieved documents and provide metrics for measuring gender bias in retrieval sets. They use an annotated dataset of gendered and non-gendered queries to demonstrate the use of these metrics in measuring gender bias of a result set.

Provider Referent (5 papers). Items can be associated with the gender of item providers or content creators (music artists, book authors), so the gender of the providers or creators is often assigned to the items [1, 20, 47]. For example, Ferraro et al. [25] identify gender bias of artists in music recommendations and propose a progressive re-ranking method that achieves improved gender balance of musical creators in recommendation systems.

Annotator Referent (1 paper). This type of paper refers to the gender of the annotators where their act of annotation is significant (compared to if they serve as test users). In the single paper in this category [79], the authors collected annotators' gender information to develop noise-aware sentiment classification models and illustrate the possible effect that demographic attributes may have on an annotator's response.

4.2.2 Gender Determination. During the coding process, we identified several ways with which authors determined the gender of the referent(s). The majority of papers (68) involved one method of gender determination, but five papers used two.

Self-identification (53 papers). In these papers, gender is determined with self-declarations of gender identity. In some cases, users declare their own gender (among other demographic attributes) while participating in a study or while using a system [7, 37, 79]; in others, authors use publicly available datasets that provide demographic data where it can be assumed that gender was self-declared [53, 62, 73].

Annotators (16 papers). In this work, human annotators assign gender for users, providers or subjects [3, 31, 70]. In [20], the authors use a dataset where the gender of book authors was annotated by library professionals.

Inferred (7 papers). In these papers, gender is interpreted from item content, users' personal information, interaction behavior or with the help of annotators. We identified papers that use users or providers name, voice, and images for inferring gender [2, 28, 43]. For example, Mukherjee et al. [47] use a gender identification tool that infer users' gender from their username and country of origin.

4.2.3 *Categories of Gender.*

Binary (63 papers). This group of papers considered gender as a binary variable where they categorized gender into men and women. This is regardless of referent or determination type. These papers also do not acknowledge that gender is non-binary [2, 7, 25, 47, 50, 60, 76].

Acknowledgement of non-binary gender, or the use of a third gender category (10 papers). The other ten papers consider the concept of gender beyond binary categorization. In six of them, the gender categories were extended to include unisex, mix-gender, and non-gendered groups. For instance, in [22], the authors considered unisex and mix-gender categories along with men's and women's categories to predict buyer's size preference in e-commerce. The remaining four papers acknowledged the limitations of representing gender as a binary construct but continued to do so in any case. For example, in [20], the authors use a binary gender variable to assess the results of collaborative filtering methods in book ratings and recommendations with respect to the gender of content creators, but include discussion of the negative effects and consequences of representing gender as binary.

Notably, none of these papers provided classifications which affirmed non-binary gender identities. This is distinct from papers which provide a "non-gendered" categorization, such as "unisex" or "other", as noted from the examples given in the prior paragraph. We discuss positive examples of affirming non-binary gender identities in the discussion.

4.2.4 *Bias and Fairness.*

With the rise in the interest of bias, fairness, and ethics in machine learning systems, and the development of new venues such as FAccT/FATML, a concomitant rise has been seen in the interest in the information access space. We coded for whether the papers dealt with issues of bias or fairness in IR systems. Of the 73 coded papers, nearly one-third (24) were concerned with bias or fairness.

4.2.5 *Purposes and Uses of Gender.*

We used an inductive coding method to assess the goal of using a gender variable. Inductive coding is typically used in grounded theory methodology [11] in which one does not presume a set of categories on some type of text, such as an interview transcript; we wanted to understand the types of goals directly from the literature instead of imposing our assumptions on it. In this case, we focused on the paper overall, rather than doing line-by-line codings.

By "goal", we refer to the intention or technical achievement attempted by the method with respect to the gender variable. This is often, but not always, distinct from the goal of the paper itself. As an example, a paper which attempts to develop a state-of-the-art collaborative filtering recommender system with demographic data as a goal may integrate a gender variable as part of a vector of demographic features. In this case, the goal would be *Gender Personalization.*

There may also be the cases in which the gender variable is used towards some other, broader end. For instance, a paper which attempts to show how errors of demographic inference get propagated in a fair ranking system would be characterized as "auditing system behavior," but not "gender prediction."

We developed ten distinct purposes of gender. The majority of papers (57) were labeled with one code but a handful (16) were coded with two or three codes. The ten purposes are outlined below.

User Study or Survey (31 papers). In this group of papers, users are asked to participate in a user study or are respondents in a survey where they assess model outcomes and provide feedback on a subjective aspect of a system. User responses are analyzed for measurements of perceived usability (user perception, user behavior, user knowledge retention). In this case, gender is often collected as a salient feature among other demographic features (age, location). For instance, in [50], the authors provided participants with a set of questions pertaining to a gender-biased result set of images to measure their perceived bias and search engine objectivity. In their assessment, the authors collected demographic information including gender, and determined measurements of two types of sexism detected in users in order to analyze the effect of a user's sexist biases on user perception of gender bias in image retrieval.

Gender Personalization (21 papers). In this group of papers, the authors use gender as part of a user profile to personalize recommendations. For instance, in [12], the authors utilized user-specific information (gender, age, social status) to improve musical artist recommendations and to assess long-term music interests of users.

Audit System Behavior (20 papers). In this genre of papers, the authors evaluate the behavior and outcomes of an existing model or framework and offer recommendations regarding functionality and/or fairness based on analysis results. Gender is highlighted among other demographic features both in the datasets used and when assessing results for fairness. For instance, in [56] the authors generated a dataset of non-gendered queries as input for several neural ranking models and measured the resulting gender bias.

Gender Prediction (7 papers). In these papers, the authors infer a gender variable from existing data instances and typically use them towards some other system end, such as improving the personalized recommendations. For instance, in [68], the authors utilized a deep-learning collaborative filtering approach to better predict size and fit of users within an e-commerce platform. To address the issue of data sparsity on user-item interactions, their model learned latent representations and implicit features of users (age, gender).

Protect Gender Variable (3 papers). In this group of papers, the main focus is privacy protection around a set of demographic variables, of which gender is highlighted. The authors often first simulated the system or model's behavior to illustrate privacy violations and/or data leakage. To counteract the issue, the authors then proposed and demonstrated an adversarial method designed

to mitigate privacy leakage and provide better protection for users' sensitive attributes, namely gender. For instance, in [39], the authors demonstrated the relative ease with which user behavioral data can be unobtrusively retrieved during web browsing via mouse cursor movements and subsequently used to predict demographic attributes (age, gender). They then provided a web browser extension that implements their proposed mitigation technique to obfuscate user demographics.

Persona Generation (3 papers). This genre of papers specifically analyzes user perceptions of profile representations derived from user data. The authors collected demographic data (age, gender) from participants and assessed the design of automatically-generated personas with respect to participant responses. In this way, gender is highlighted as a demographic point of interest in both users and user perceptions of gendered personas. In [61], for instance, the authors conducted a survey measuring user perceptions of pseudo-personas, specifically in response to pairs of identical profiles where the profile features a smiling picture versus a non-smiling picture. They found gender to be an influential attribute of generated personas, wherein variation in the gender of participants resulted in perceptual variation of the gendered personas.

Indexing Clinical Trials (2 papers). In this genre of papers, the authors evaluate query expansion and reduction techniques and work to determine optimal feature configurations to improve information retrieval within the medical field. The authors utilize a gender variable (among other demographics) to improve query results. In [1], for instance, the authors evaluated a precision medicine search engine and its functionality in retrieving scientific literature and clinical trials in which they employ four steps: an indexing step, a query reformulation step, a retrieval step, and a filtering step. In the indexing step, the authors included a gender field (among other demographic fields) to index clinical trials and used these fields to determine eligibility in the filtering step.

Gender Diversity & Inclusion (1 paper). In this body of papers, the methods involve using gender, amongst other demographic attributes, to algorithmically determine diversity and inclusion in model outputs or a UX surface. In the single example in our dataset [47], the authors offered an unsupervised summarization framework that provides a user with control over the shape and content (e.g., the gender of reviewers) of aspect-based summaries of tourist reviews on TripAdvisor.

Linguistic Gender (1 paper). This set of papers deal with how to negotiate gendered aspects of language, including pronouns, nouns, and other gendered components. In our single example [77], the authors morphologically annotated Amharic (a gendered language) for the purpose of extending the application of lexical analysis to include more languages.

Gender Interest Personalization (1 paper). In this last group of papers, they deal with dyadic gender preferences, rather than the gender of the referent themselves, which would fall under the concept of Gender Personalization. In our sole example [42], the authors focused on a dating app context where "match" suggestions depend upon the user's specific gender preferences of prospective companions.

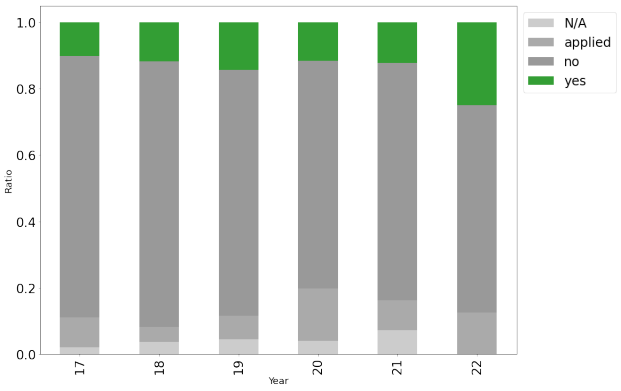

Figure 1: Breakdown of whether the paper had a gender variable by year. "N/A" is used when papers refer to phrases such as "sexuality" and not biological sex.

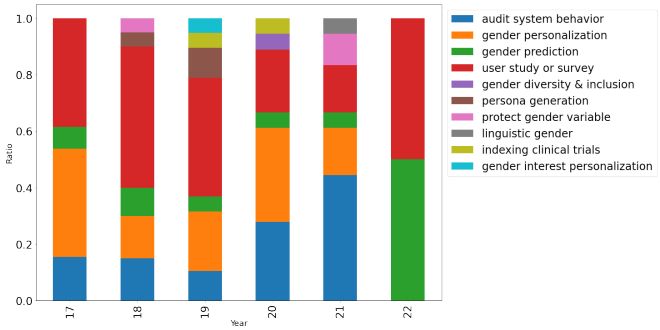

Figure 2: Goals across time

4.3 Bivariate Analysis

The prior section provided an overview of our data findings for each of the respective variables we coded for in our review of papers. In this section, we dig into some of the trends of data across time and variables.

4.3.1 Time Trends. Figure 1 shows the breakdown of our dataset by year. There has not been more of a focus on gender across time. There is a slight increase in the number of papers which mention a gender term, but about the same proportion of papers contain a gender variable from year to year. However, there are some notable changes across the goals of the use of a gender variable across time.

The goals of using a gender variable have changed across time. The top two goals ("user study or survey" and "gender personalization") are somewhat persistent across the study period, with the prior category peaking in 2018 and the latter in 2020. However, our third most prevalent category ("audit system behavior") has been steadily climbing since the beginning of the study period, with its peak in 2021.

Similarly, the use of a gender variable with the intent of assessing or testing for some kind of bias or fairness issue has risen across time, from two papers in 2017 to eight papers in 2021. In fact, in 2021, the majority of papers (8 of 15) dealt with fairness issues.

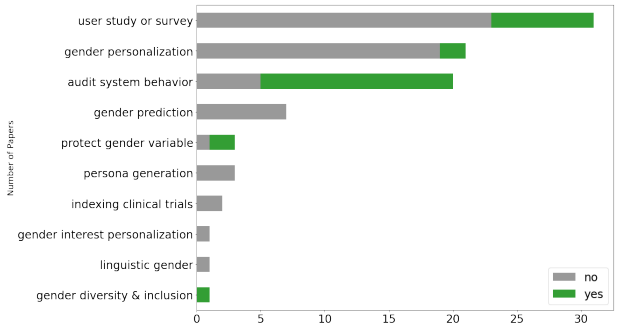

Figure 3: Breakdown of bias and fairness by goal.

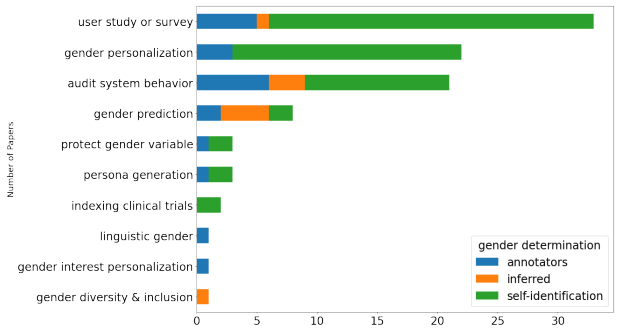

Figure 4: Gender determination by different goals of the gender variable.

4.3.2 Bias and Fairness. When looking at the evaluation of bias and fairness as it pertains to each individual goal (Figure 3), we have found papers which audit system behavior address this topic significantly more often than papers which use gender variables towards any other end. User studies and surveys address bias at the second highest rate. None of our coded papers with the goal of "gender prediction" address bias or fairness in their discussion, and only two of those papers with the goal of "gender personalization" make note of this topic. The one paper which addresses gender diversity and inclusion deals with fairness issues.

4.3.3 Gender Determination and Goals. Figure 4 shows the bivariate relationship between gender determination and paper goal. Most of the papers used "self-identification" as a gender determination. This is overwhelmingly the case for user studies (27), gender personalization (19), and auditing of system behavior (12). However, only two papers with the goal of "gender prediction" use "self-identification" as a gender determination, whereas all but two papers regarding "gender personalization" use "self-identification" over both inference and annotation. Significantly, papers which do gender prediction mostly use an inferred gender, which is not surprising, given the method. However, three papers which audit a system's behavior use inferred gender, and one uses it in the case of user studies.

4.4 Discussion

From our analysis, there are several areas worth noting with regards to the use of a gender variable. Most notably, we found no positive incorporation of non-binary genders within the papers we reviewed: that is, no papers successfully affirmed or accounted for non-binary gender identities. Although there a small portion of papers provided additional categories of gender beyond the binary male and female labels, it is important to note that the absence or neutrality of gender (as implied by "unisex" and "non-gendered" classifications) is not synonymous with non-binary gender identities. Over time, it appears that discussion, or, at the very least, acknowledgement of gender as non-binary has increased, but the successful utilization of a non-binary gender variable has yet to be made.

Secondly, there has been more awareness in fairness-oriented uses of gender variables in this research community, and it has gone up over time. Although there appears to be more of an effort on this front with goals like "audit system behavior", it remains that papers with the goal of "gender personalization" and "gender prediction" fail to properly analyze the implications of their findings or model behavior in reference to gender bias and fairness. However, it may be the case that these two types of goals are antagonistic or fundamentally at odds with fairness and ethics, as suggested by Keyes [36] and Scheuerman et al. [65].

Third, the most frequent goal of using a gender variable is as input to an analysis in a user study or survey. This suggests that these authors are studying how differently gendered individuals respond to particular systems, which may be an encouraging result. More troubling, however, is the frequency at which systems attempt to personalize results based on gender. This itself makes major assumptions about what individuals may prefer, based on a gender variable, rather than on user preferences. We discuss alternative practices of personalization below. A more heartening development, however, is that auditing of system behavior has increased over the past five years, and that most of these studies do this with some kind of fairness evaluation in mind.

Lastly, across all papers, gender self-identification is the norm, rather than the exception. Self-identification is the most ethical manner of collecting gender data, although the exact method of doing so is still an area of discussion and research, as noted in our literature review above. In a small number of cases, however, gender is inferred or labeled by third-party annotators. Third-party evaluations, either by crowdworkers, paper authors, or machines, may perpetuate gender stereotypes or be another vector of misgendering. When users self-declare their own gender identities within a dataset, they are less likely to be misgendered by a system or model using that data than when human annotators or systems infer gender identities from data traces, such as product selection, names, face images, or texts that the individual writes. Self-declaration of gender, however, does not foreclose the possibility of misgendering, because much self-identification data are collected with only binary gender categories built into the systems which collect these data in the first place.

5 RECOMMENDATIONS

Researchers and practitioners need to proceed with care in dealing with gender in computational research. Depending on the goal, use,

and determination of gender, both the research process and findings of such research may be harmful. This harm may be direct, as when a system misgenders a person, or it may be indirect by handling gender in a reasonable way on its own but when combined with other downstream components causes harm. In this section, we provide some high-level recommendations and guidelines about using gender information in research on information access systems. We are not providing definite rules of using gender in computational research; rather, we are providing recommendations that researchers and practitioners can consider to avoid inappropriate use of gender in their work. We also expect future work to build on these guidelines as both understanding and technical possibilities evolve.

5.1 When to Use Gender?

Researchers should first determine whether it is appropriate to use gender in the first place. For some applications, contexts, or goals, using gender in some way may be beneficial; for others, it may just not be useful; and in a number of cases it is likely actually harmful.

Auditing system performance, particularly for fairness and equity concerns (the goal of 20 of 73 papers) seems a relatively positive use of gender. Its purpose is to identify and mitigate gendered harms the system may inflict or reproduce, and the results are usually only made visible in aggregate (so errors in gender determination are rolled up in statistical aggregates, rather than present in a table of genders of individual people, although public datasets to support such audits do include individual-level gender annotations). For example, Ramos and Boratto [55] examined systems that rank people and may have reputational implications to ensure that the resulting reputation is independent of gender. Care is needed, though, in order not to undermine the fairness or equity goal: work that aims to improve fairness but only does so within a binary gender construct, for example, may reinforce discrimination against non-binary people. Moreover, audits of system behavior that infers gender on individuals may reproduce harm by guaranteeing that a system works only for individuals who conform to stereotypical gender presentations or expressions. Lastly, this work may be used to diversify information access systems (e.g. [46]), but the same caveats for doing so via gender inferrence remains.

Overall, we advise against personalization based on gender as a goal or component of a system (the goal of 21 papers). Such personalization inherently depends on stereotypes about peoples' interests and capabilities, either existing stereotypes derived from societal assumptions or new ones derived from data. This contradicts the premise of online personalization based on extensive user profiles, as implemented in collaborative filtering, that we can personalize to a user's particular needs and tastes rather than relying on unpersonalized or group-based assumptions. As Riedl and Konstan argued [58], recommender systems should "box products, not people." The literature we have surveyed has not made a compelling case for gender-based personalization, but rather assumes that it is a reasonable thing to do or does it because it has been done before. There is also reason to be suspicious of using gender for personalization even in cold-start scenarios before individual user feedback is available: because the feedback from which personalized systems learn is not entirely exogenous, but is partly a response to the system's previous outputs [10], the system may learn future "data-driven" stereotypes not from organic user interactions but from its own initial assumptions. That is, if initial recommendations are derived from erroneous gender stereotype assumptions, data from the resulting interactions may reinforce those assumptions not because they are an accurate model of user interests, but because the user would have clicked on any comparable recommendation. Further study is needed to identify whether and to what extent this is happening, but it is a risk that should be taken seriously.

Lastly, following critical work on automated gender recognition [36, 65], we also advise against gender prediction in information access systems (the goal of 7 papers). Many of the papers we find in our data focused on gender prediction aim to make that determination from user behavior, such as written internet text [59] or more esoteric data such as spatial trajectories [69]. However, similar to our warning against gender personalization above, these predictions may perpetuate gender stereotypes and re-entrench them by making those determinations based on data instances which bear no relationship to gender, and will most likely misrepresent individuals who are transgender or gender non-conforming.

5.2 How to Use Gender?

If it is appropriate to consider using gender in some way, actually operationalizing and applying it requires additional careful consideration. In this section, we focus on more ethical goals of using gender and ethical strategies of gender determination.

Our first recommendation is to use an inclusive concept of gender to the extent possible. Restricting work to a male/female gender binary limits its applicability and reproduces exclusion of gender minorities. Data selection is the first obvious application of this principle, but it goes beyond simply the data; for example, while Ekstrand and Kluver [18] (expanding on [20]) acknowledged non-binary gender identities as valid and an important limitation, the metric and resulting statistical method they employed cannot be applied to non-binary attributes. Even when only binary data is available, we advise against methods that cannot be applied outside of binary contexts, so that the analysis can be updated if and when more inclusive data is located or produced [54].

Examples of inclusive gender data and analyses are rare, but the TREC Fair Ranking track and dataset [19] does use non-binary gender identities for bibliographic Wikipedia articles where available. The appendix of the track description [19] provides full details of the gender attribute, but they started with 20+ gender identities from Wikidata, collapsed transgender identities (treating trans men as men and trans women as women), and folding remaining gender identities into a third category; this resulted in "male", "female", "third" ("nonbinary" in 2022), and "unknown." This has the benefits of reducing combinatorial explosion and the number of groups with very few representatives, making the encoding more computationally practical. One downside of this approach is that it may obscure discrimination against binary transgender people specifically.

Our second recommendation is to document precisely how gender labels were obtained, whatever schema they use; prior work demonstrates that many datasets do not justify where they obtain the data nor the schema of data labels [67]. This recommendation

applies to both data obtained from existing sources, including public datasets, and new datasets created for particular projects. Such documentation should be reported in relevant publications and can also be a part of dataset documentation such as a datasheet [27]. This document should document the schema used, the source of the data (such as self-identification or expert annotation), the construct of gender recorded (e.g. gender identity versus gender expression), and the principles used to determine gender when it is not self-identification. In specifying the schema, the documentation should also describe the options given to respondents, and if various instruments or interfaces limited options to the male/female gender binary. When working with existing data sets, such information may not be immediately obvious, but researchers and practitioners alike should perform due diligence to understand how gender was collected and recorded before working with the data. Documenting this information can serve as a community benefit for other researchers who seek to build on their results and/or work with the same data. When data is obtained from an intermediary, both the intermediary and the intermediary's source of gender data should be identified. In many of the papers we coded, the paper was not explicit about the source of gender data, and we had to infer the source from context, background assumptions, or other resources.

Our third recommendation is to consider greater gender diversity in one's data sample, especially when conducting small-n qualitative studies or user studies in which gender may be a significant factor for understanding results. We found that only 10 of our 73 papers which used a gender variable acknowledged non-binary gender, or provided a third option. None, however, positively affirmed a non-binary gender option. Therefore, it would be highly advisable that non-binary people are explicitly recruited for studies in which gender could be a key variable for both the auditing of a system, or for user studies which evaluate a system.

Finally, when constructing new data sets for either research or application purposes, we recommend collection and curation that is thoughtful and respectful towards different gender identities, as well as taking into account that there is a danger in collecting demographic information in and of itself, as such information may make reductionist assumptions about identity [32], or be used in a way that violates privacy [26]. Self-identification is the best way to obtain gender data, as it most fully respects individual autonomy and self-determination, and it should be obtained through inclusive means. The HCI Gender Guidelines [66] provides guidance for how to design gender-inclusive survey fields to obtain gender information from respondents. Expert annotation can be legitimate, but should be done in a way that respects peoples' right to self-identify, along with their right to be excluded entirely. The Program for Cooperative Cataloging established a task force to produce recommendations for how to record the gender identities of book authors in library name authority records [5], whose report provides explicit guidance about the type of inferences that should or should not be used when recording author information (when an author does not state their gender identity, the recommendations allow inference from clear indications in sources close to the author, such as the choice of gendered pronouns in an author's own biography, but not from names or photographs). The relevant data field is also explicitly defined as recording an author's gender identity [40].

5.3 Research Needed

Our systematic review and the recommendations we draw from it and relevant literature and guidance in adjacent fields are by no means the last word on the use and misuse of gender in information access. Further research is needed to identify and assess the various impacts of use-of-gender decisions. There are also open practical challenges: for example, while there is important work on measuring fairness beyond binaries [54, 78], it is not easy to deal computationally with rich notions of gender that may be multidimensional, combinatorially large, and have categories with relatively few members. When it is appropriate to use gender — for example, in audits for discrimination — the details of how to ethically, respectfully, and practically collect, store, document, analyze, and present rich notions of gender remain to be worked out.

There is also space to carry out similar analyses to understand how gender is being used in other fields such as natural language processing or data mining, and to document the use of gender in deployed industrial systems that are not yet described in the public research literature.

6 CONCLUSION

Gender is a complex and multifaceted construct that is often connected with important aspects of a person's identity. A review of published literature reveals a variety of goals for which gender is employed. Pursuing gender equity in the effects of information access systems is an important goal, but this needs to be done thoughtfully and in a manner that respects the rights and identities of the people involved. Sometimes, gender should not be used; in other cases, it should be used but with due care and attention to the complexity of gender. This also needs to be accompanied with clear discussions of what, precisely, has been done, why, and limitations that arise from the chosen approach.

Our aim with this paper has been to provide an understanding of the current state of research practice and pointers to further reading to understand gender as it is currently understood, to serve as a foundation for robust, rigorous, and respectful investigations of how information access systems can avoid reproducing gender-related harms and can effectively serve users, content creators, and information subjects of all genders.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. 17-51278.

REFERENCES

[1] Maristella Agosti, Giorgio Maria Di Nunzio, and Stefano Marchesin. 2019. An Analysis of Query Reformulation Techniques for Precision Medicine. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, 973–976. https://doi.org/10.1145/3331184.3331289

[2] Ali Ahmadvand, Harshita Sahijwani, and Eugene Agichtein. 2020. Would You Like to Talk About Sports Now? Towards Contextual Topic Suggestion for Open-Domain Conversational Agents. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, 83–92. https://doi.org/10.1145/3343413.3377974

[3] Pinar Barlas, Styliani Kleanthous, Kyriakos Kyriakou, and Jahna Otterbacher. 2019. What Makes an Image Tagger Fair?. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, 95–103. https://doi.org/10.1145/3320435.3320442

[4] Manash Pratim Barman, Amit Awekar, and Sambhav Kothari. 2019. Decoding The Style And Bias of Song Lyrics. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, 1165–1168. https://doi.org/10.1145/3331184.3331363

[5] Amber Billey, Matthew Haugen, John Hostage, Nancy Sack, and Adam L Schiff. 2016. *Report of the PCC Ad Hoc Task Group on Gender in Name Authority Records*. Technical Report. Program for Cooperative Cataloging. https://www.loc.gov/aba/pcc/documents/Gender_375%20field_RecommendationReport.pdf

[6] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (July 2016). arXiv:1607.06520 [cs.CL] https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[7] Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, and Katya Pertsova. 2018. How to Evaluate Humorous Response Generation, Seriously?. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, 225–228. https://doi.org/10.1145/3176349.3176879

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81 (2018), 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline

[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically From Language Corpora Contain Human-Like Biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[10] Allison J B Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, 224–232. https://doi.org/10.1145/3240323.3240370

[11] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE. https://play.google.com/store/books/details?id=2ThdBAAAQBAJ

[12] Zhiyong Cheng, Jialie Shen, Liqiang Nie, Tat-Seng Chua, and Mohan Kankanhalli. 2017. Exploring User-Specific Information in Music Retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, 655–664. https://doi.org/10.1145/3077136.3080772

[13] Matthew Crain. 2018. The Limits of Transparency: Data Brokers and Commodification. *New Media & Society* 20, 1 (Jan. 2018), 88–104. https://doi.org/10.1177/1461444816657096

[14] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, 2083–2102. https://doi.org/10.1145/3531146.3534627

[15] Mark Diaz and Emily Denton. 2021. A Dataset Exploration Case Study with Know Your Data. https://ai.googleblog.com/2021/08/a-dataset-exploration-case-study-with.html. https://ai.googleblog.com/2021/08/a-dataset-exploration-case-study-with.html Accessed: 2022-10-18.

[16] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism*. MIT press.

[17] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177. https://doi.org/10.1561/1500000079

[18] Michael D Ekstrand and Daniel Kluver. 2021. Exploring Author Gender in Book Rating and Recommendation. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 377–420. https://doi.org/10.1007/s11257-020-09284-2

[19] Michael D Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the TREC 2021 Fair Ranking Track. In *The Thirtieth Text Retrieval Conference (TREC 2021) Proceedings*. https://trec.nist.gov/pubs/trec30/papers/Overview-F.pdf

[20] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver British Columbia Canada). Association for Computing Machinery, 242–250. https://doi.org/10.1145/3240323.3240373

[21] Avriel Epps-Darling, Romain Takeo Bouyer, and Henriette Cramer. 2020. Artist Gender Representation in Music Streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 248–254. https://program.ismir2020.net/poster_2-11.html

[22] Yotam Eshel, Or Levi, Haggai Roitman, and Alexander Nus. 2021. PreSizE: Predicting Size in E-Commerce Using Transformers. (May 2021). arXiv:2105.01564 [cs.IR] http://arxiv.org/abs/2105.01564

[23] Darrell Etherington. 2016. Baidu and KFC's New Smart Restaurant Suggests What to Order Based on Your Face. *TechCrunch* (Dec. 2016). https://techcrunch.com/2016/12/23/baidu-and-kfcs-new-smart-restaurant-suggests-what-to-order-based-on-your-face/

[24] Anne Fausto-Sterling. 2000. *Sexing the Body: Gender Politics and the Construction of Sexuality*. Basic Books. https://play.google.com/store/books/details?id=c3lhYfZzIXkC

[25] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, 249–254. https://doi.org/10.1145/3406522.3446033

[26] Samantha Floreani. 2021. Privacy and Gender: What to Ask, When and Why. https://www.salingerprivacy.com.au/2021/09/13/privacy-and-gender/. https://www.salingerprivacy.com.au/2021/09/13/privacy-and-gender/ Accessed: 2023-1-9.

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. 2018. Datasheets for Datasets. (March 2018). arXiv:1803.09010 [cs.DB] http://arxiv.org/abs/1803.09010

[28] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, 1033–1043. https://doi.org/10.1145/3404835.3462850

[29] Kevin Guyan. 2022. *Queer Data: Using Gender, Sex and Sexuality Data for Action*. Bloomsbury Academic.

[30] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, 8. https://doi.org/10.1145/3173574.3173582

[31] Kyungsik Han, Yonggeol Jo, Youngseung Jeon, Bogoan Kim, Junho Song, and Sang-Wook Kim. 2018. Photos Don't Have Me, But How Do You Know Me? Analyzing and Predicting Users on Instagram. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, 251–256. https://doi.org/10.1145/3213586.3225232

[32] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 2020)*. Association for Computing Machinery, 501–512. https://doi.org/10.1145/3351095.3372826

[33] Seyyed Hadi Hashemi and Jaap Kamps. 2017. Skip or Stay. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17* (Oslo, Norway). ACM Press. https://doi.org/10.1145/3020165.3022160

[34] Amy Hawkins. 2017. KFC China is Using Facial Recognition Tech to Serve Customers - But Are They Buying It? *The Guardian* (Jan. 2017). https://www.theguardian.com/technology/2017/jan/11/china-beijing-first-smart-restaurant-kfc-facial-recognition

[35] Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, 23–28. https://doi.org/10.1145/3213586.3226206

[36] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. https://doi.org/10.1145/3274357

[37] Akiva Kleinerman, Ariel Rosenfeld, and Sarit Kraus. 2018. Providing Explanations for Recommendations in Reciprocal Environments. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, 22–30. https://doi.org/10.1145/3240323.3240362

[38] Y Koren, R Bell, and C Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (Aug. 2009), 30–37. https://doi.org/10.1109/MC.2009.263

[39] Luis A Leiva, Ioannis Arapakis, and Costas Iordanou. 2021. My Mouse, My Rules: Privacy Issues of Behavioral User Profiling via Mouse Tracking. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, 51–61. https://doi.org/10.1145/3406522.3446011

[40] Library of Congress Network Development. 2022. 375 - Gender (R). https://www.loc.gov/marc/authority/ad375.html. https://www.loc.gov/marc/authority/ad375.html Accessed: 2022-10-17.

[41] Devon Magliozzi, Aliya Saperstein, and Laurel Westbrook. 2016. Scaling Up: Representing Gender Diversity in Survey Research. *Socius* 2 (Jan. 2016), 2378023116664352. https://doi.org/10.1177/2378023116664352

[42] Rahul Makhijani, Shreya Chakrabarti, Dale Struble, and Yi Liu. 2019. LORE: A Large-Scale Offer Recommendation Engine With Eligibility and Capacity Constraints. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, 160–168. https://doi.org/10.1145/3298689.3347027

[43] Jennifer Marlow and Jason Wiese. 2017. Surveying User Reactions to Recommendations Based on Inferences Made by Face Detection Technology. In *Proceedings of*

the Eleventh ACM Conference on Recommender Systems (Como Italy). Association for Computing Machinery. https://doi.org/10.1145/3109859.3109875

[44] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 626–633. https://doi.org/10.1145/3041021.3054197

[45] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–23. https://doi.org/10.1145/3449100

[46] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, 117–123. https://doi.org/10.1145/3375627.3375832

[47] Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read What You Need: Controllable Aspect-Based Opinion Summarization of Tourist Reviews. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 1825–1828. https://doi.org/10.1145/3397271.3401269

[48] Reza Nasirigerdeh, Reihaneh Torkzadehmahani, Jan Baumbach, and David B Blumenthal. 2021. On the Privacy of Federated Pipelines. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, 1975–1979. https://doi.org/10.1145/3404835.3462996

[49] Safiya Umoja Noble. 2018. *Algorithms of Oppression*. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

[50] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, 933–936. https://doi.org/10.1145/3209978.3210094

[51] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents* (Glasgow, Scotland, UK) *(SMUC '11)*. Association for Computing Machinery, 37–44. https://doi.org/10.1145/2065023.2065035

[52] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much Ado About Gender (CHIIR 2023) Codebook and Data. https://doi.org/10.5281/zenodo.7521838

[53] Yanru Qu, Bohui Fang, Weinan Zhang, Ruiming Tang, Minzhe Niu, Huifeng Guo, Yong Yu, and Xiuqiang He. 2018. Product-Based Neural Networks for User Response Prediction Over Multi-Field Categorical Data. *ACM Transactions on Information and System Security* 37, 1 (Oct. 2018), 1–35. https://doi.org/10.1145/3233770

[54] Amifa Raj and Michael D Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press. https://doi.org/10.1145/3477495.3532018

[55] Guilherme Ramos and Ludovico Boratto. 2020. Reputation (In)dependence in Ranking Systems: Demographics Influence Over Output Disparities. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 2061–2064. https://doi.org/10.1145/3397271.3401278

[56] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 2065–2068. https://doi.org/10.1145/3397271.3401280

[57] Elaine Rich. 1979. User Modeling via Stereotypes. *Cognitive Science* 3, 4 (Oct. 1979), 329–354. http://www.sciencedirect.com/science/article/B6W48-4FWF9GC-9/2/f924f793eb153d455893e8d39982ef45

[58] John Riedl and Joseph Konstan. 2002. *Word of Mouse*. Warner Books.

[59] Ohad Rozen, Joel Oren, and Ariel Raviv. 2021. Predicting User Demography and Device from News Comments. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, 1995–1999. https://doi.org/10.1145/3404835.3463024

[60] Joni Salminen, Bernard J Jansen, Jisun An, Soon-Gyo Jung, Lene Nielsen, and Haewoon Kwak. 2018. Fixation and Confusion: Investigating Eye-tracking Participants' Exposure to Information in Personas. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, 110–119. https://doi.org/10.1145/3176349.3176391

[61] Joni Salminen, Soon-Gyo Jung, João M Santos, and Bernard J Jansen. 2019. The Effect of Smiling Pictures on Perceptions of Personas. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, 75–79. https://doi.org/10.1145/3314183.3324973

[62] Pablo Sánchez and Alejandro Bellogín. 2019. Attribute-Based Evaluation for Recommender Systems: Incorporating User and Item Attributes in Evaluation Metrics. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, 378–382. https://doi.org/10.1145/3298689.3347049

[63] Veronica Sanz. 2017. No Way Out of the Binary: A Critical History of the Scientific Production of Sex. *Signs: Journal of Women in Culture and Society* 43, 1 (Sept. 2017), 1–27. https://doi.org/10.1086/692517

[64] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-Essentialization: Gender in Automated Facial Analysis as Extended Colonial Project. *Big Data & Society* 8, 2 (July 2021), 20539517211053712. https://doi.org/10.1177/20539517211053712

[65] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–33. https://doi.org/10.1145/3359246

[66] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI Gender Guidelines. https://www.morgan-klaus.com/gender-guidelines.html. https://www.morgan-klaus.com/gender-guidelines.html Accessed: 2020-5-21.

[67] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–35. https://doi.org/10.1145/3392866

[68] Abdul-Saboor Sheikh, Romain Guigoures, Evgenii Koriagin, Yuen King Ho, Reza Shirvany, Roland Vollgraf, and Urs Bergmann. 2019. A Deep Learning System for Predicting Size and Fit in Fashion E-Commerce. (July 2019). arXiv:1907.09844 [cs.LG] http://arxiv.org/abs/1907.09844

[69] Adir Solomon, Ariel Bar, Chen Yanai, Bracha Shapira, and Lior Rokach. 2018. Predict Demographic Information Using Word2vec on Spatial Trajectories. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, 331–339. https://doi.org/10.1145/3209219.3209224

[70] Xuemeng Song, Xiang Wang, Liqiang Nie, Xiangnan He, Zhumin Chen, and Wei Liu. 2018. A Personal Privacy Preserving Framework: I Let You Know Who Can See What. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, 295–304. https://doi.org/10.1145/3209978.3209995

[71] Maryam Tavakol. 2020. Fair Classification with Counterfactual Learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 2073–2076. https://doi.org/10.1145/3397271.3401291

[72] The GenIUSS Group. 2014. *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-based Surveys*. Williams Institute, UCLA School of Law. https://play.google.com/store/books/details?id=T1sDogEACAAJ

[73] Yuan Tian, Ke Zhou, Mounia Lalmas, Yiqun Liu, and Dan Pelleg. 2020. Cohort Modeling Based App Category Usage Prediction. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, 248–256. https://doi.org/10.1145/3340631.3394849

[74] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. arXiv:2109.05433 [cs.CV] http://arxiv.org/abs/2109.05433

[75] Laurel Westbrook and Aliya Saperstein. 2015. New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys. *Gender & Society* 29, 4 (2015), 534–560. https://doi.org/10.1177/0891243215584758

[76] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. 2020. Joint Item Recommendation and Attribute Inference: An Adaptive Graph Convolutional Network Approach. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 679–688. https://doi.org/10.1145/3397271.3401144

[77] Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2021. Morphologically Annotated Amharic Text Corpora. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, 2349–2355. https://doi.org/10.1145/3404835.3463237

[78] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with Multiple Protected Groups. *Information processing & management* 59, 1 (Jan. 2022), 102707. https://doi.org/10.1016/j.ipm.2021.102707

[79] Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. 2019. Learning from Multi-annotator Data: A Noise-Aware Classification Framework. *ACM Transactions on Information and System Security* 37, 2 (Feb. 2019), 1–28. https://doi.org/10.1145/3309543