

Investigating the Capabilities and Limitations of Machine Learning for Identifying Bias in English Language Data with Information and Heritage Professionals

Lucy Havens
 hello@lucyhavens.com
 University of Edinburgh
 Edinburgh, UK

Melissa Terras
 m.terras@ed.ac.uk
 University of Edinburgh
 Edinburgh, UK

Benjamin Bach
 benjamin.bach@inria.fr
 Inria
 Bordeaux, FR
 University of Edinburgh
 Edinburgh, UK

Beatrice Alex
 b.alex@ed.ac.uk
 University of Edinburgh
 Edinburgh, UK

Abstract

Despite numerous efforts to mitigate their biases, ML systems continue to harm already-marginalized people. While predominant ML approaches assume bias can be removed and fair models can be created, we show that these are not always possible, nor desirable, goals. We reframe the problem of ML bias by creating models to identify biased language, drawing attention to a dataset's biases rather than trying to remove them. Then, through a workshop, we evaluated the models for a specific use case: workflows of information and heritage professionals. Our findings demonstrate the limitations of ML for identifying bias due to its contextual nature, the way in which approaches to mitigating it can simultaneously privilege and oppress different communities, and its inevitability. We demonstrate the need to expand ML approaches to bias and fairness, providing a mixed-methods approach to investigating the feasibility of removing bias or achieving fairness in a given ML use case.

CCS Concepts

- Human-centered computing → HCI design and evaluation methods;
- Social and professional topics → Socio-technical systems;
- Applied computing → Digital libraries and archives;
- Computing methodologies → Machine learning; Natural language processing.

Keywords

Human-Centered Machine Learning, Human-Centered AI, Gender Bias, Data Bias, Language Bias, Cultural Heritage

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '25, April 26-May 1, 2025, Yokohama, Japan

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
 ACM ISBN 979-8-4007-1394-1/25/04
<https://doi.org/10.1145/3706598.3713217>

ACM Reference Format:

Lucy Havens, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2025. Investigating the Capabilities and Limitations of Machine Learning for Identifying Bias in English Language Data with Information and Heritage Professionals. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3706598.3713217>

1 Introduction

The Gallery, Library, Archive, and Museum (GLAM) sector collects and provides access to *cultural heritage*, meaning records deemed culturally or historically significant [129, 138, 144]. These records include audio recordings, books, letters, maps, musical instruments, paintings, and social media posts, among others. To enable their discovery and use, librarians, archivists, and curators (“information and heritage professionals”) document them in *catalogs*, writing metadata descriptions that summarize the records’ cultural, economic, historical, political, social, and temporal contexts [2, 145]. These descriptions shape researchers and the general public’s understanding of the past because they determine how cultural heritage records can be discovered and provide an initial interpretation of the records [126]. Though for many years describing cultural heritage records was viewed as an objective process, since the late 19th century, information and heritage professionals have increasingly recognized description as an act of interpretation and knowledge production, often undertaken in Western nations within the context of imperialism and colonialism [41, 132]. Consequently, descriptions of cultural heritage records in GLAM institutions’ catalogs, as well as the records themselves, reflect and reinforce power relationships within and between nations. To counteract the marginalization of certain communities of people that catalog descriptions have reflected and reinforced, information and heritage professionals have begun dedicating resources to reviewing the descriptions for biased language, a process often referred to as *critical cataloging* [9].

To support critical cataloging, especially for large catalogs with millions of entries created over several centuries, Machine Learning (ML) methods increasingly seem promising. Recent advances in ML have led to higher performing and more readily adoptable large

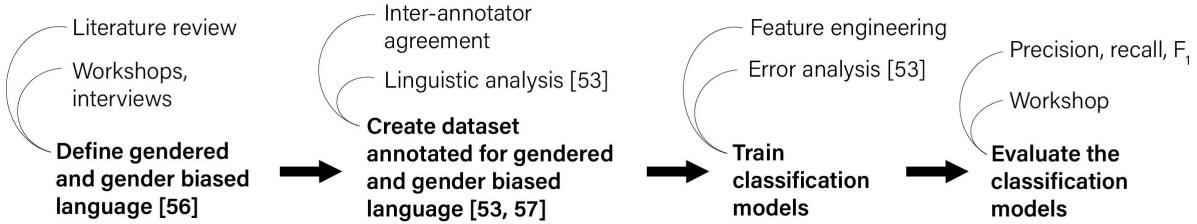


Figure 1: A flow chart of our four-step research process, with each step in boldface and the main tasks for the steps radiating out and above each step. This paper summarizes the entire research process while focusing on the last step, “Evaluate the classification models,” and the substep “Workshop,” through which we qualitatively evaluate the models (§5). Previous publications, as referenced in the flow chart and throughout this paper, provide more detail on the steps for defining gendered and gender biased language [56], creating the annotated dataset [57], and training and quantitatively evaluating the classification models [53].

language models, such as OpenAI’s GPT-4 [107]. However, ML research and practice focus on use cases that are not relevant to information and heritage professionals in the GLAM sector for three reasons. Firstly, predominant ML approaches to bias aim to remove bias from data or models, eliminating the possibility of analyzing bias as a historical phenomenon with ramifications for present-day society [56, 129]. Secondly, despite many ML publications claiming the wide generalizability of their models [117], GLAM publications report that state-of-the-art ML models’ training datasets are too different from GLAM catalog data to make the models suitable to GLAM use cases [96]. Thirdly, as we demonstrate in this paper, the types of language bias considered in ML approaches are not comprehensive enough for the GLAM sector because they typically rely on the definition and connotation of specific words, aiming to replace terms considered discriminatory with today’s preferred terms, such as replacing “transsexual” with “transgender.” There are many types of bias that are more complicated to detect, however, and that have implications for both information and heritage professionals’ efforts to manage bias in descriptions of cultural heritage, and for ML researchers and practitioners’ approach to data and model bias (§2).

Given the rapidly growing adoption of ML systems by the public, such as Google’s Gemini,¹ and the growing integration of ML models into products, such as Microsoft’s Copilot,² underestimation of the biases in ML systems’ underlying data threatens to magnify social injustice at an unprecedented scale. In this paper, we combine Human-Computer Interaction (HCI) and ML methods to investigate how bias manifests in data, specifically gender bias in the language of a British GLAM catalog’s descriptions. We aim to support information and heritage professionals in managing bias in GLAM collections, and to expand ML researchers and practitioners’ understandings of the varieties of bias to consider when creating and deploying ML systems. Additionally, we aim to bridge the GLAM and ML communities, encouraging ML researchers and practitioners to draw upon information and heritage professionals’ approaches and practices, because the GLAM sector has a long history of managing bias in collections at the community, national,

and international scales at which ML systems are being deployed. To that end, we asked the following research questions:

- **RQ1:** How do information and heritage professionals conceptualize bias? To what extent does this align with how ML researchers and practitioners conceptualize bias?
- **RQ2:** How can the identification of gender biased language with ML models inform and impact information and heritage professionals’ understanding of gender biased language and management of bias in their institutions’ catalogs and collections?

First, we created ML models to identify gender bias in a GLAM catalog’s metadata descriptions (§4). Then, using a participatory action research approach [134] grounded in critical discourse analysis [43] and intersectional feminism [62], we ran a workshop with information and heritage professionals to evaluate the ML models (§5). We conducted content analysis on the workshop participants’ discussion, finding that (a) information and heritage professionals do conceptualize bias differently than ML researchers and practitioners, viewing bias as contextual, dynamic, and inevitable; and (b) ML models can expand information and heritage professionals’ understanding of biased language and bias in their institutions’ collections, though (c) ML models are limited in their ability to identify data bias comprehensively and reliably due to the contextual, dynamic, and inevitable nature of bias. In addition to supporting information and heritage professionals’ management of bias, our findings have implications for HCI researchers and practitioners working with ML systems. Observations from the workshop (§6) highlight manifestations of bias that require information and heritage professionals’ domain expertise to identify and manage, reinforcing the necessity of human-centered research methods in ML approaches to bias and fairness. Our findings also have implications for ML researchers and practitioners, surfacing complexities of data bias that have been overlooked and demonstrating that, in certain use cases, removing bias and creating fair ML systems may not be possible nor desirable. Our mixed-methods evaluation of ML models and their underlying data provides an approach to investigating the feasibility of removing bias or achieving fairness in a given ML use case.

¹gemini.google.com

²copilot.microsoft.com

2 Background

Existing work in the GLAM sector acknowledges how societal power relationships shape what GLAM institutions collect, how those collections are described, and how these processes shape cultural and historical narratives that produce and perpetuate social biases [28, 41]. For example, information and heritage professionals have critiqued how these processes have enacted privilege and oppression based on gender [106], sexuality [1], race [44], and colonialism [132]. Existing GLAM work also acknowledges the power that information and heritage professionals have to reject and reshape narratives that portray inaccurate and harmful views of particular communities [24, 25]. Aiming to empower historically marginalized communities, information and heritage professionals have been reviewing and revising the catalog metadata descriptions that summarize and contextualize their GLAM institutions' collections. Such anti-oppressive, critical cataloging work has been largely manual [9, 136], requiring GLAM institutions to divide their already limited resources between describing new collections and revisiting existing descriptions.

ML affords efficiency and scalability that appeals to information and heritage professionals because it would advance the profession's primary goal of providing access to information [2]. The ability to more quickly describe heritage could help address the backlog of items that have not been or were only minimally cataloged [50, 140]. The Library of Congress, for instance, adds approximately 10,000 new items to its collections every day and employs approximately 3,200 people, only a fraction of whom are catalogers responsible for describing incoming items [92, 93]. Though experiments with automated metadata generation are underway [70, 87], GLAM institutions still rely on information and heritage professionals' manual review and description of items to ensure their discoverability for the public. Existing research identifies three factors limiting ML methods and models' applicability to GLAM use cases: (1) a domain mismatch between state-of-the-art ML models' training data and GLAM catalog data [30, 55, 96], (2) a skills gap between typical information and heritage professionals' education and training and the data and computer science education and training needed to create and use ML models [53, 98, 137], and, most importantly for this paper, (3) a mismatch between conceptualizations of bias and fairness between information and heritage professionals and ML researchers and practitioners [27, 116]. Research demonstrating the inefficacy of ML approaches to minimizing bias (§3.1) justify information and heritage professionals' cautious approach to adopting pre-trained ML models.

Due to the complexity of social biases, which results from multifaceted and ever-changing societal power relationships [33, 34, 62, 146], we focus our work on gender bias, working with British English text from a Scottish Archive's catalog metadata descriptions. We adopt our previously published definition of gender biased language as: "*language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their gender identity; and privileging other people through words or phrases that favor their gender identity*" [57, p. 108]. We are interested in how information and heritage professionals can provide insight on implicit ways that gender bias manifests in language, as opposed to

the explicit forms of biased language more widely studied in ML: primarily terminology that is derogatory or stereotypical for content moderation and machine translation use cases [26, 124]. ML publications on gender bias have reported omissions of women in training data [21], stereotypical representations of men and women in models [103], and omissions and misrepresentations of people of trans and gender diverse identities in models [78]. Research on sexism in computing has demonstrated the socio-technical nature of bias, with gender biased social structures resulting in datasets that encode gender biases and models that have gender biases engineered into them [60, 61, 113]. These patterns in ML reflect patterns in GLAM: cultural heritage records about women and people of trans and gender diverse identities have been under-collected [59, 128] and, when collected, under-documented and misrepresented [106], perpetuating gender stereotypes and under-acknowledging the contributions of these gender communities [7]. Such omission and misrepresentation has wide-reaching consequences for women and people of trans and gender diverse identities, affecting their experiences in everything from transportation to medical care to work opportunities [15, 42, 113].

The challenge with addressing gender biased language is that it is subjective and contextual. The same words or phrases may be interpreted differently by different people and may be harmful in one context but not another, such as varying attitudes towards the term "queer" and changes in the term's connotation over time in Western cultures [19, 125, 128]. Consider the gender bias in this description of Dr. Susan Binnie Anderson (1898-1975):

As a young married woman with two small children—Kathleen and Margaret—she found it difficult to keep her medical career going [...] She twice tried to return to medicine full-time [sic] but this just wasn't to be.³

There is no derogatory terminology or explicitly stereotypical word choice in the description. Yet, one could argue the description communicates an implicit assumption that the state of being young, married, a woman, and a mother negates the possibility of a medical career. Patriarchal societal structures during Dr. Anderson's lifetime may have caused the difficulty she faced, but the description does not consider this. On the other hand, the description does not provide information about Dr. Anderson's husband, who may have also found it difficult to keep his career going being young, married, and a father, in which case the difficulty of maintaining a career would be implicitly attributed to the challenges of parenthood. Some people may identify gender bias in this description while others may not. This implicit type of bias and the uncertainty surrounding it is understudied in ML research and, with the growing adoption of generative ML models for search, summarization, writing, and decision making, of increasing concern [65, 85, 97].

3 Related Work

3.1 Bias in ML

ML approaches to data and model bias predominantly view bias as a technical flaw that needs to be mitigated in order to create fair ML systems (for summaries, see [29, 67, 101, 114, 133]). Approaches to creating unbiased, fair ML systems include efforts to

³archives.collections.ed.ac.uk/repositories/2/resources/546

measure bias in datasets [5, 74] and models [71, 72], create balanced datasets [143, 148] and create or fine-tune models to treat different demographic groups fairly [81, 123]. Several scholars have noted that predominant approaches to ML bias and fairness rely on quantitative analyses and mathematical representations, failing to account for the societal structures that cause and perpetuate social biases [6, 64, 75, 100]. The work of Rodolfa *et al.* [120], Ciora *et al.* [26], and Samorani *et al.* [123] illustrate how considering the distribution of power in the societal context in which an ML system will be deployed can inform the system's design. However, the focus of these scholars' work was creating ML systems that do not behave in a biased manner. More research is needed on use cases where biases are inevitable or full automation is not desirable, such as in healthcare, criminal justice, and the GLAM sector.

Despite the proliferation of ML bias mitigation methods put forth, reliable, generalizable mitigation methods have yet to be found. Research inspecting the efficacy of bias mitigation approaches has yielded inconsistent results, with some findings indicating successful mitigation [72] and others indicating no conclusive impact [46, 86, 108]. Most concerning are findings that indicate existing approaches have amplified bias [65] and become new causes of harm to the marginalized communities that the approaches were intended to help [29, 36]. While recently proposed methods of matched guise probing [65] and mechanistic interpretability [77, 91] offer promising approaches for more effectively detecting bias, understanding how biases are engineered into models and comprehensively testing for them are ongoing challenges. This may be due to manifestations of bias that are not well-understood, such as the covert forms of sexism Ciora *et al.* investigate for machine translation in Turkish [26] and covert forms of racism that Hofmann *et al.* found in large language models [65]. This may also be due to overconfidence in the generalizability of datasets, models, benchmarks, and metrics [6, 112, 117, 141]. Our work aims not only to better understand the limitations of ML approaches to bias, but also to demonstrate how to test for those limitations.

To improve ML approaches to bias and fairness, scholars have called for greater interdisciplinary research and context-specific approaches [13, 29]. This has led to ML research that draws upon sociolinguistics [23], feminism [79], legal and political concepts [3, 75], and human-centered research methods common to the HCI community [85, 99, 102, 142], among others. HCI scholars have grown the fields of human-centered ML and eXplainable Artificial Intelligence (XAI), using surveys, interviews, and participatory research methods to study ML systems in specific contexts [90, 127, 135]. This work has distinguished the needs of different end users in different usage contexts [35, 89] and shed insight on users' perceptions of ML systems, including perceptions of bias and fairness, and the implications of these perceptions for users' workflows [66, 104, 122, 147]. Still, research is needed to investigate how an ML system's end users conceptualize bias beyond the confines of an ML system; that is the focus of this paper.

Although the GLAM sector has a longer history of managing bias than the ML community, few ML researchers or practitioners have drawn upon GLAM approaches for guidance on working with data. Jo and Gebru encourage ML researchers and practitioners to look to archival and library cataloging practices for guidance on dataset documentation [73]. Thylstrup *et al.* use archives as a

metaphor for big data, exploring the questions, challenges, and opportunities that arise when approaching large-scale datasets as archival collections [139]. Crawford and Paglen note that information and heritage professionals have long recognized the political implications of classification in a comparison of ImageNet's classification scheme with a historical version of the Library of Congress Classification scheme [32]. Yet there are also GLAM approaches with more practical applicability to ML.

3.2 Bias in GLAM

GLAM approaches to bias offer insight on why efforts to mitigate bias in ML systems continue to fall short. Given the predominance of manual data work in the GLAM sector, information and heritage professionals are particularly attune to the implications of inequitable societal power relationships on creating, collecting, organizing, analyzing, and using data [17, 24, 41, 59, 109, 132]. They have looked to critical theories to develop new practices that minimize harms from biases in GLAM catalogs and the collections they describe. Nonetheless, information and heritage professionals' conceptualizations of bias have not been characterized in a way that is easily translatable to training and evaluating ML models. In this paper, by evaluating ML models with information and heritage professionals (§5), we are able characterize bias in a manner that draws upon GLAM domain expertise and is translatable to ML system creation (§7).

Critical theories originate in social justice movements and interdisciplinary research, emphasizing context-specific approaches that consider power and social and cultural reproduction [83, 88]. Critical heritage studies scholar Smith draws on *critical discourse analysis* [129], which views the meaning of language as a result of the words used and the context in which language is produced and received [20, 43]. The former is referred to as the "internal relations" of a text; the latter, the "external relations" [43]. Smith draws on critical discourse analysis to expand conceptualizations of heritage beyond the predominant Western idea of large, tangible artifacts, defining heritage as a process of creating, sharing, using, and recreating knowledge [129]. We extend her framework to data, considering how internal and external relations of an ML model's training data affect the biases of a model. By starting with biased language in the context of ML systems, existing ML research focuses on the internal relations of data, such as explicitly derogatory language [121] or pornographic images [12].⁴ We expand upon this work by starting from human understandings of social biases, or the external relations of a dataset, facilitating a discussion about bias with our training data's domain experts: information and heritage professionals (§5). This enables us investigate which types of biases ML models cannot reliably identify or avoid reinforcing.

⁴There is ML research on bias that considers external relations such as how skews in research participants' demographic characteristics can bias research findings [94] and how gender biases in language relate to the person producing, receiving, or being described in text [39]. However, the manifestation of bias being studied in this area of work is about agreements and disagreements between research participants or data labels. Moreover, in this area, researchers often view disagreeing labels as an indication of different interpretations of text [110]. We position our work among research that studies how bias manifests in language and allows for the possibility that disagreeing labels may indicate different opinions about text that has, in fact, been interpreted to mean the same thing [4, 47]. See [55] for a broader consideration of bias in GLAM catalogs and collections.

Table 1: The coding taxonomy of our training dataset's source data, reproduced with authors' permission [57]. In the far right column, coded text is boldfaced and, for the four middle rows, co-referents determining the coding of a name are italicized.

Code	Code Category	Explanation	Example Description (Coded Text in Boldface)
Generalization	Linguistic	Gender-specific term used to refer to a group that could include more than one gender	His classes included Anatomy, Practical Anatomy, Midwifery and Diseases of Women.
Gendered Role	Linguistic	Grammatically- or lexically-gendered word	New map of Scotland for Ladies Needlework.
Gendered Pronoun	Linguistic	Feminine, masculine, or non-binary (neo-) pronouns	He obtained surgical qualifications in 1873.
Feminine	Person Name	Person referred to with a grammatically or lexically feminine term	Although Yolanda Sonnabend shared the copyright of this book with Waddington, <i>she</i> was not credited as an illustrator.
Masculine	Person Name	Person referred to with a grammatically or lexically masculine term	Martin Luther , the <i>man</i> and <i>his</i> work.
Non-binary	Person Name	Person referred to with a grammatically or lexically non-binary term	Francis McDonald went to the University of Edinburgh where <i>they</i> studied law.
Unknown	Person Name	Person referred to without a grammatically or lexically gendered term	Testimonials in favor of Niecks .
Occupation	Contextual	Job title	He became a surgeon with the Indian Medical Service.
Omission	Contextual	Exclusion of a person's identity or contribution	This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version.
Stereotype	Contextual	Reductive or derogatory expectation of a person's behaviors or preferences	Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more.

Considering data's external relations aligns with *feminism*, a critical theory that rejects the idea of a universal truth, instead viewing knowledge as situated [51, 52]. Information and heritage professionals have drawn upon feminism for managing bias in GLAM collections [24, 25, 106]. Havens draws on intersectional feminism [33, 34, 62] to inform GLAM description practices, including critical cataloging [54]. She proposes using Hill Collins' matrix of domination [62] when analyzing the external relations that influence the meaning of catalog metadata descriptions and how biases manifest in their language. The matrix of domination consists of four interacting domains through which power is exercised in society: the *structural* domain organizes oppression and privilege (e.g., a government body), the *disciplinary* domain manages oppression and privilege (e.g., a court), the *hegemonic* domain justifies oppression and privilege (e.g., a museum), and the *interpersonal* domain is where oppression and privilege are experienced (e.g., by a person). The matrix of domination provides a structure around which to conceptualize power relationships between individuals, communities, and organizations that create social biases, oppressing some and privileging others. We used critical discourse analysis and intersectional feminism as lenses to guide our workshop analysis, through which we sought to understand why ML methods alone are insufficient for understanding how bias manifests in data (§6).

4 Machine Learning

We use ML to reframe the problem of bias. Our use of ML aligns with concepts of "sandcastling" [63] and "doing design otherwise" [40], which emphasize the process of building something. Hinrichs *et al.* use the term "sandcastle" to describe data visualizations built for digital humanities projects that "facilitate a collaborative discourse about the means through which information is developed, collected, and represented" [63, p. i86]. DiSalvo describes "doing design otherwise" as the use of design methods to interrogate problems rather than create solutions to problems, engaging in social and political efforts in the process [40]. For both sandcastling and doing design otherwise, what is built is situated, fragile, and incomplete; the research value is in the questions and insights gleaned through the critical thinking and collaborative, cross-disciplinary analysis that the building process surfaces. The fragile and incomplete nature of what is built encourages cross-disciplinary participation, discussion, and critique [40, 63]. As such, the built thing has a purpose similar to that of a technology probe [68] or boundary object [130]. The ML models we develop and use in this paper function more as sandcastles than as algorithmic innovations, enabling us to facilitate a discussion with GLAM domain experts to surface questions and insights about manifestations of gender bias that ML models can and cannot identify (§5). Moreover, given growing evidence of

how ML innovations are reinforcing the power of the world's most privileged people rather than having a democratizing impact on global society [10, 31, 49, 58, 60], our use of ML upends this trend by heightening the visibility of data bias. We use ML to call attention to inequitable gender-based power relationships, informing and motivating efforts to empower women and people of trans and gender diverse identities.

4.1 Training Dataset

We created our ML models' training dataset from our previously published, human-coded datasets of descriptions from the archival catalog of the University of Edinburgh's Heritage Collections [57]. The catalog is organized according to the General International Standard for Archival Description (ISAD(G)) [69]. The dataset's descriptions came from four metadata fields: "Title," "Scope and Contents," "Biographical / Historical," and "Processing Information." The descriptions document heritage in a range of formats (e.g., letters, photographs, degree certificates) covering a variety of topics (e.g., religion, research, teaching, architecture). The coding taxonomy has ten types of gendered and gender biased language: ***Gendered Pronoun***, ***Gendered Role***, ***Generalization***, ***Feminine***, ***Masculine***, ***Non-binary***, ***Unknown***, ***Occupation***, ***Omission***, and ***Stereotype*** (Table 1). The dataset contains 11,888 descriptions, totaling 399,957 words across 24,474 sentences, with a total of 55,260 codes. Only 9 of the codes were applied; ***Non-binary*** is not present in the dataset. A.1 details our preprocessing work to transform the dataset for model training and testing. For linguistic analysis of the coded data, see [53, p. 86–93].

4.2 Model Design, Training, and Evaluation

Given our interest in supporting information and heritage professionals' understanding of gender bias in a specific GLAM catalog, we created text classification models in a supervised, traditional ML setup. This ensured that the gender biases the models identified originated in our training dataset. Though deep learning models are considered state-of-the-art, they pose a risk to the validity of our research: biases from a pre-trained, or foundation, model's initial training dataset would be injected into our customized version of the model [46, 84, 131]. ML researchers and practitioners have yet to develop an approach to distinguishing biases in a pre-trained model's training dataset from biases in the dataset used to customize that model. Since we aimed to identify gender biased language in our training dataset only, we did not use pre-trained deep learning models, because their classification of the descriptions would have been influenced by biased language in other data. Given the unique task of our models, namely, classifying text as ***Omission*** or ***Stereotype*** based on the definitions in Table 1, the models we report in this paper provide a baseline against which future ML experiments can be compared. In the remainder of this section, we briefly summarize the training and testing process for our classification models; §A.2 further details the process for reproducibility.

We created three types of ML models. The Linguistic Classifier (LC) is a multilabel token classification model that codes grammatically or lexically gendered words, classifying text as ***Gendered Pronoun***, ***Gendered Role***, and ***Generalization***. The Person Name

and Occupation Classifier (PNOC) is a multiclass sequence classification model coding phrases, sentences, or paragraphs based on the grammatical or lexical gender of a person's co-referents (see the right-most column of Table 1 for examples), classifying text as ***Feminine***, ***Masculine***, and ***Unknown***;⁵ and coding job titles, classifying text as ***Occupation***. The Omission and Stereotype Classifier (OSC) is a multilabel document classification model that codes descriptions as gender biased, classifying text as ***Omission*** and ***Stereotype***.

In addition to creating these models to interrogate conceptualizations of gender biased language, we were interested in investigating correlations between grammatically and lexically gendered language and gender biased language. For instance, would identifying the presence of gendered roles in a description (e.g., "Lady," "father," "Mrs.") help a classifier determine whether or not a description communicates a gender stereotype? To investigate this, we trained baseline versions (meaning no feature engineering) of the LC, PNOC, and OSC, and then we sequentially combined these classifiers in cascades, using a previous classifier's predictions as features for the subsequent classifier in a cascade, resulting in three feature-engineered versions of the OSC. The cascades are:

- **Cascade 1: LC > PNOC > OSC:** We ran the LC and then passed its predictions (***Gendered Pronoun***, ***Gendered Role***, and ***Generalization*** codes) to a PNOC as features. Then, we ran that PNOC and passed its predictions (***Feminine***, ***Masculine***, ***Unknown***, and ***Occupation*** codes) and the LC's predictions to an OSC as features.
- **Cascade 2: LC > OSC:** We ran the LC and then passed its predictions (***Gendered Pronoun***, ***Gendered Role***, and ***Generalization*** codes) to an OSC as features.
- **Cascade 3: PNOC > OSC:** We ran a PNOC and then passed its predictions (***Feminine***, ***Masculine***, ***Unknown***, and ***Occupation*** codes) to an OSC as features.

By comparing the performance scores of the baseline and feature-engineered OSCs, we can evaluate whether including gendered language codes as features improves an OSC's ability to classify text as gender biased.

⁵The ***Non-binary*** code does not appear in the training data so it could not be used in the models.

Table 2: Macro performance scores for each Omission and Stereotype Classifier (OSC), the average of the human coders' Inter-Annotator Agreement (IAA) scores, and the average of the five human coders' agreement scores with the training dataset. Human coder scores are reproduced with permission [57].

	Precision	Recall	F ₁
Baseline OSC	0.899	0.643	0.747
Cascade 1 OSC	0.896	0.644	0.747
Cascade 2 OSC	0.889	0.667	0.760
Cascade 3 OSC	0.888	0.669	0.761
Coders' IAA	0.485	0.420	0.427
Coders vs. Train	0.996	0.662	0.786

Table 3: Comparing human coders and text classification models’ performance for *Omission* and *Stereotype*. The “Coders’ IAA” column reports the average of the IAA scores between the human coders [57]. The “Coders vs. Train” column displays the average of the human coders’ agreement scores with the training dataset. The “OSC” columns report the performance scores of Omission and Stereotype Classifiers. Scores are calculated loosely, meaning two codes are considered to be in agreement if applied to overlapping or exactly matching text spans. The highest scores per label are in bold. Human coder scores are reproduced with permission [57].

	Coders’ IAA			Coders vs. Train			Baseline OSC		
Label	Precision	Recall	F ₁	Precision	Recall	F ₁	Precision	Recall	F ₁
Omission	0.525	0.400	0.429	0.996	0.631	0.757	0.859	0.539	0.663
Stereotype	0.446	0.441	0.426	0.995	0.694	0.815	0.939	0.746	0.832
	Cascade 1 OSC			Cascade 2 OSC			Cascade 3 OSC		
Label	Precision	Recall	F ₁	Precision	Recall	F ₁	Precision	Recall	F ₁
Omission	0.853	0.546	0.666	0.848	0.563	0.676	0.845	0.571	0.682
Stereotype	0.940	0.741	0.829	0.930	0.771	0.843	0.932	0.766	0.841

All models were trained and tested using a modified five-fold cross-validation approach (see §A.2 for detail). We evaluated the models using precision, recall, and F₁ scores, optimizing for F₁ score (§4.3).

4.3 Model Performance

In this section we summarize the results of our ML experiments most relevant to the workshop (§5): the classification of gender biased language in descriptions with the *Omission* and *Stereotype* codes. §A.3 reports on model performance for the remaining codes in the taxonomy. For further error analysis and discussion of the classification results, see [53, p. 149–166].

The OSCs’ F₁ scores suggest that coding tokens as grammatically or lexically gendered improves a model’s ability to classify gender biased language because the feature-engineered OSCs in cascades 1-3 outperform the baseline OSC (Table 2). The OSCs’ performance scores exceed the human coders’ IAA overall (across both codes) and per *Omission* and *Stereotype* code. The OSCs’ recall and F₁ scores also exceed the human coders’ agreement with the training dataset for the *Stereotype* code, but not overall, due to the OSCs’ lower performance classifying text with the *Omission* code. Given that the training dataset has over twice as many *Omission* codes (4,032) than *Stereotype* codes (1,601), the greater difficulty classifying descriptions with *Omission* than with *Stereotype* is surprising. However, the OSCs’ lower scores for *Omission* aligns with human coders’ scores, which were also lower for *Omission*. These results suggest that there may be greater variation in the language coded as *Omission* than as *Stereotype*.

Though the OSCs’ performance scores are low relative to state-of-the-art classification models that rely on deep learning methods, the low human coder IAA scores indicate that classifying descriptions as *Omission* and *Stereotype* is a very difficult task. Our classifiers are the first ML models built to identify manifestations of gender biased language in a GLAM catalog’s metadata descriptions, so they provide a baseline upon which future work can build (§7.3). The models’ performance demonstrates the viability of training ML

models on GLAM catalog metadata, which has been questioned in previous work [96].

Our model design pushes back against predominant approaches to creating and evaluating ML systems. Model performance metrics typically do not consider bias or fairness, and the identification of bias in a model does not affect its status as “state-of-the-art;” consider the continued widespread use of GPT-4 despite Hofmann *et al.*’s findings of racial bias in the model [65]. Given the misrepresentations and omissions that result from biases in ML systems, we argue that performance evaluations should, as a standard, include measures of bias and fairness. By creating ML models trained to classify gender biased language, we unavoidably tie bias measurement to performance evaluation.

The next section reports our human-centered ML evaluation: a discussion of our models’ training data and output classifications in a workshop. Our workshop participants (§5) are information and heritage professionals who work in the University of Edinburgh’s Heritage Collections, so their knowledge of the collections represented in our training dataset enable them to provide an informed evaluation of the ML models’ classifications, comparing and contrasting their interpretation of a description’s gender bias with the models’ interpretation. By creating ML models to classify descriptions rather than manually coding descriptions for gender biased language ourselves, we avoid coding descriptions in a manner that emphasizes only the biases of which we as authors are most cognizant (the classifiers were trained on data that aggregates five expert annotators’ coding [57]). Additionally, creating the models enabled us to quickly classify more descriptions than we could have coded manually, providing the data needed for workshop activity 2. Our models-as-sandcastles thus have two purposes: (a) for information and heritage professionals, illustrating how ML could enable more efficient, large-scale reviews of GLAM catalog descriptions for bias; (b) for us as researchers, motivating cross-disciplinary discussion that provides insights on the capabilities and limitations of ML for identifying biased language in a real-world use case. As

we describe in §7, these insights have implications for all ML researchers and practitioners, providing guidance on how to improve dataset curation, data coding, and model design processes.

5 Workshop

5.1 Setup and Procedure

We facilitated an open-format workshop to evaluate the ML models with 10 information and heritage professionals. We wrote pre-defined questions to guide participants' discussion during the workshop in the manner of a semi-structured interview [118], which we refined during a pilot workshop with three doctoral students and one postdoctoral researcher in digital humanities and information science. The workshop was approved by the ethics board of the School of Informatics at the University of Edinburgh (reference 2019/81479). The workshop discussion was audio-recorded with participants' consent using a single microphone connected to Zoom. We corrected the Zoom-generated transcript of the workshop manually and shared the transcript with participants to ensure its accuracy prior to analyzing the discussion.

The workshop consisted of an introduction, two activities, and a wrap-up question, lasting two hours total. We began the workshop by explaining the supervised, traditional ML method we used to create text classification models, which we had used to classify descriptions from the University of Edinburgh's archival catalog (*i.e.*, the participants' GLAM institution's catalog) for gender bias. Then we began Activity 1 (A1), which focused on investigating RQ1: *How do information and heritage professionals in the GLAM sector conceptualize bias? To what extent does this align with how ML researchers and practitioners conceptualize bias?* We gave each participant a worksheet with the codes that the models were trained to classify with on one side (Figure 2) and three coded descriptions on the other side (Figure 3). We selected the coded descriptions from our models' training data because we aimed the discussion during A1 to focus on the interpretation of the descriptions and meaning of the codes, facilitating a human-centered evaluation of the models' design. To start the discussion, we asked participants:

- Q1.1: Do you agree or disagree with, or are you unsure about, the labels on the descriptions? Why?
- Q1.2: How would you use this information?
- Q1.3 What information is missing that you would need to support your work?

A1 lasted one hour. We took a 10-minute break and then began the second activity.

Activity 2 (A2) focused on RQ2: *How can the identification of gender biased language with ML models inform and impact information and heritage professionals' understanding of gender biased language and management of bias in their institutions' cultural heritage collections?* We gave each participant a worksheet (figures 4 and 5) with tables and charts displaying quantities of gendered and gender biased language in the University of Edinburgh's archival catalog, and a chart displaying performance measures for the Linguistic Classifier. We calculated these quantities based on the output codes of Cascade 2's LC and OSC and the output codes of the baseline PNOC because these were the best-performing versions of our models (see A.3). We presented participants with these quantities to

facilitate a human-centered evaluation of our models' outputs and performance. To start the discussion, we asked participants:

- Q2.1: What do you understand from the information on the worksheet? What questions do you have?
- Q2.2: How would you use this information?
- Q2.3: What information is missing that you would need to support your work?

A2 lasted 40 minutes.

We concluded the workshop by asking participants one wrap-up question:

- Q3.1: Is there any information on the worksheets you would want to share with visitors (*i.e.*, researchers, the general public)?

The wrap-up activity lasted five minutes. We sent participants a questionnaire consisting of 18 multiple choice and short answer questions after the workshop.

5.2 Participants

We recruited participants through an email to 14 information and heritage professionals working on the University of Edinburgh's Heritage Collections teams, 10 of whom participated. According to our post-workshop questionnaire, participants' job titles can be broadly categorized as archivist (six participants), curator (one participant), librarian (one participant), and manager (one participant). They described their job responsibilities as including acquisition, cataloging, curation, processing, development, management, preservation, interpretation, and facilitation of discovery and access of collections; teaching and giving training about collections; overseeing colleagues and volunteers; responding to inquiries about collections; communicating with donors and stakeholders of collections; participating in international communities of practice; and research. They reported encountering biases in GLAM catalogs and collections related to gender, sexuality, faith, race and ethnicity, nationality, language, accent, and disability. Participants' GLAM work experience ranged from six years to over 10 years. They described their gender as "female," "cisgender female," and "female-ish."

5.3 Analysis

We performed content analysis on the workshop discussion using a grounded theory approach [45, 82, 119], performing open coding on the transcript of the discussion. Our approach to open coding was rooted in intersectional feminism [33, 62] and critical discourse analysis [20, 43]. We also referenced participants' responses to the post-workshop questionnaire and their notes on the worksheets during our content analysis. We report the results of this analysis as "observations" (§6), which we shared with workshop participants prior to publishing to provide an opportunity for clarification and feedback. From these observations we identified three areas of value the models held for information and heritage professionals' management of bias, and three characteristics of data bias for ML researchers and practitioners' consideration (§7).

Gender and Gender Biased Language Classifier Review

Taxonomy of Gendered and Gender Biased Language

Labels are grouped into three categories; the labels are the bolded, lettered list items.

1. Person Name – labels applied based on the grammatical gender of terms referring to a person in the description the name appears in
 - a. Unknown
 - b. Feminine
 - c. Masculine
 - d. Nonbinary – *not used during the annotation process*
2. Linguistic
 - a. Gendered Pronoun
 - b. Gendered Role – e.g., mother, Sir, boy, Queen
 - c. Generalization – assumption made about a people's gender identity (e.g., a computer scientist always carries his laptop to work)
3. Contextual
 - a. Occupation – job title, e.g., baker, author, Professor of...
 - b. Stereotype*
 - c. Omission*

*All labels except 3b and 3c are applied on individual words; 3b and 3c are applied to entire descriptions.

Notes on the Dataset

Descriptions were extracted from four metadata fields of the Archives catalogue: Title, Scope and Contents, Biographical / Historical & Processing Information.

The data viewed in this workshop is based on the first 20% of the Archives catalogue as it was when the data was extracted in April 2020.

Figure 2: The front of the worksheet participants were given for the workshop's first activity displaying the coding taxonomy that outlines how the ML models were trained to classify descriptions.

Example 1

Broadcast of the service commemorating [sic] the centenary of the death of **Unknown** **Henry Duncan** at which
Masculine **John Baillie** was **Descriptive** **Possessor** **He** gave a biographical talk on **Unknown** **Duncan**'s life and work.

Example 2

Mixture of press cuttings covering many subjects including articles on Dietrich Bonhoeffer, ... Second World War,
 housekeeping tips and matters of general interest to **Florence Jewel Baillie**.

Example 3

Correspondence and related items, relating to the attendance [sic] of **John Baillie** and
Possessor **Ruler** **his wife** at the coronation of **Unknown** **Elizabeth II**.

Figure 3: The back of the worksheet participants were given for the workshop's first activity with three metadata descriptions from the University of Edinburgh's archival catalog, coded according to the taxonomy (Table 1). Together the three descriptions provide examples of how to apply every code in the taxonomy.

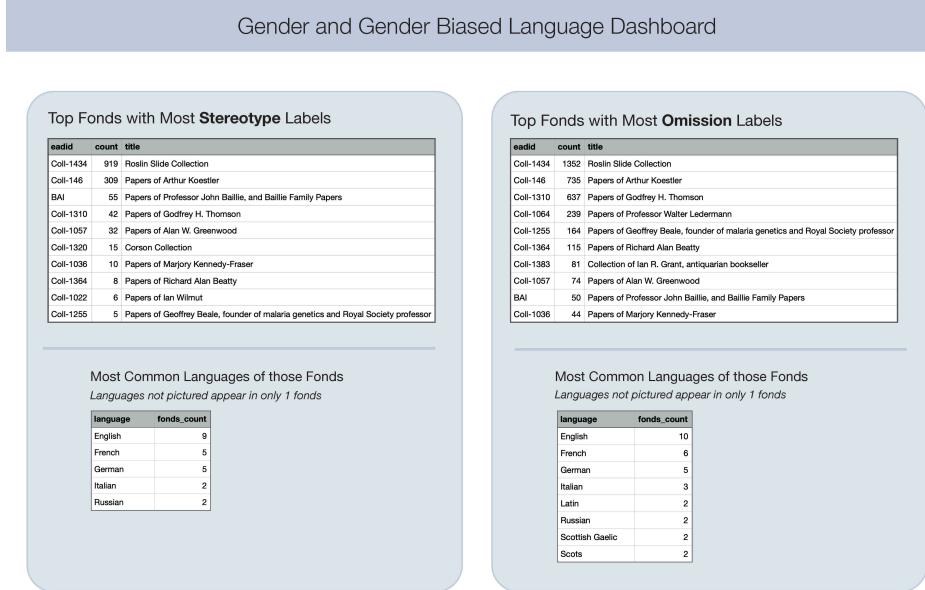


Figure 4: The front of the worksheet participants were given for the workshop's second activity. The tables at the top show total quantities of *Stereotype* and *Omission* codes in the collections ("fonds") with the highest total counts of those codes. The tables at the bottom show the most common languages of the archival records in the collections listed in the tables above.



Figure 5: The back of the worksheet participants were given for the workshop's second activity displaying two charts with quantities of taxonomy codes across the entire training dataset, and one chart visualizing the performance of a classifier as represented by true positive, false positive, true negative, and false negative counts.

6 Results

Below we summarize 15 observations (O1-O15) that emerged from our content analysis, grouping the observations by our research questions. Participants are referred to anonymously as P1, P2, etc.

6.1 RQ1: Conceptualizations of Bias

Regarding observations relevant to RQ1, participants' comments communicated a conceptualization of bias that is context-dependent, changing based on who is represented in the data and who is looking at the data, including when and where those people are located and those people's lived experiences.

O1: Concern with losing gender information. At the beginning of the workshop, participants' discussion communicated a concern about the codes highlighting language that should be removed or changed. P6 explained how, due to the historical and ongoing marginalization of women in many societies around the world, often a woman is only identified in cultural heritage records with a title and last name (e.g., "Mrs. MacDonald"), so, "*if you lose 'Mrs.' then you lose all of their identity rather than enhancing their identity.*" P8 gave an example of how the identity of a woman, who was depicted in a painting with her husband, was able to be discovered thanks to the more extensive documentation of her husband and the description of the woman as a "Mrs." Through discovering the woman's identity, P8 was able to learn that the wife brought significant funds to she and her husband's marriage that enabled the couple to develop a substantial art collection. The inclusion of grammatically and lexically gendered terms can thus provide clues to guide research into the contributions of women, supporting efforts to address the lack of documentation about women's roles and contributions throughout history [7, 48, 59].

Participants also discussed the value of identifying people by gender even if that gender is assumed. Regarding the application of the **Unknown** code, P7 stated, "*I feel kind of weird, though, with it, in that now I can't say what their genders are [...] how can we say who they are if we can't assume that?*" Participants explained how describing people with gendered terminology can be useful even if an assumption was made about a person's gender, because that assumption reflects how the person was perceived in society during their lifetime. During the workshop, in response to participants' concern about losing information with the removal of gendered terminology, we clarified that the taxonomy's codes were not intended to highlight language for removal or change, but rather were intended to make potential gender biases more easily identifiable.

O2: Gender bias in the codes' application. Regarding the application of the **Occupation** code, participants suggested that it could have been applied more broadly than only to job titles. Looking at examples 2 and 3 in Figure 3, P10 asked why "housekeeping" did not have an **Occupation** code and then said, "*And it's the same thing, you know, with 'coronation,' hinting at her [Queen Elizabeth II's] occupation.* Despite the intention of the codes to push back against such gender biases [57], their application reinforced occupational gender biases and the under-documentation of women's historical contributions. Women have faced greater barriers to entering the workforce and obtaining job titles, but this does not mean women have not worked.

O3: Subjectivity of bias. Differences in the interpretations of the codes, as well as an awareness of the subjectivity of bias and language, came through in participants' comments and questions. **Stereotype** yielded the greatest debate. Participants asked questions about how **Stereotype** had been defined and requested additional examples of text that would be classified with this code. P5 spoke of the contextual nature of a stereotype, stating, "*It's changed by people, by culture, by things like this, so it seems really hard to define what it is.*" P7 expressed concern about the judgment implied with the **Stereotype** code in the second description (Figure 3), stating, "*You could be doing her a huge injustice if her life's mission was housekeeping.*" This description prompted the greatest discussion about stereotypes. P8 requested clarification about whether the **Stereotype** code was in fact applied due to the association of a woman with "housekeeping." P3 responded:

"I thought it wasn't just the housekeeping but that a woman's interests just get summarized into 'general interests.' [...] It doesn't just say, 'Florence Jewel Bailie was an avid researcher of Dietrich Bonhoeffer, the Second World War, housekeeping, etc.' It just says here are some news cuttings on some things that she read about. Which kind of makes a judgment. That's what I thought the stereotype was in this instance."

O4: Value of dates to contextualize data. Participants expressed the importance of including dates where available because dates situate a description's language in the historical context in which the language was produced. Regarding potential biases of descriptions in the "Title" catalog metadata field, P3 said, "*we probably inherit that [...] so it'd be good to be able to isolate language that's been created recently by an archivist.*" P1 and P3 spoke of the value of dates alongside a cataloging system that would maintain the version history of a description, which is not provided in current cataloging systems' functionality. P1 explained, "*it's about dating descriptions and not overwriting them.*"

O5: Sense of responsibility to past, present, and future stakeholders of catalogued cultural heritage. Participants communicated a sense of responsibility to historical people and communities represented in cultural heritage records, to present-day descendants of those people and members of those communities, and to future visitors who will search GLAM catalogs for records of those people and communities. Reflecting on the role of GLAM catalogs and GLAM description practices, P9 commented:

"I think it's sort of socially really important [...] because so many people are going to look it [the catalog] up, students or members of the public, and refer to that as a kind of non-biased factual resource. So it's different from writing something where you assess your positionality and make that clear to the reader [...] it's a big responsibility to get right as well."

Participants expressed concern about the challenge of navigating this responsibility when different stakeholders hold conflicting preferences or points of view. Recalling cataloging work with cultural heritage records in which a person referred to themselves as a "Highland traveler," a term today considered derogatory, P6 recounted how a present-day member of this community had contacted the

University of Edinburgh's Archive to request that the catalog description be changed to use today's accepted term, "Gypsy Roma Traveler." P6 explained the conflicting responsibilities information and heritage professionals must grapple with in such situations: "*So you then get into layers of, well who's say is it? Is it the current community or is it the person whose own voice it is?*" P4 gave an example of this conflict regarding gender, explaining that historical members of the trans community used terms to describe themselves that are considered offensive today. Nonetheless, P4 viewed these terms as important to include in catalog descriptions of cultural heritage records because such terminology is, "*something that's an example of trans history or gender non-conformity.*"

In response to this discussion, P7 asked, "*What do you do when [...] in 20 years' time, something you created is deemed offensive?*" P2 added that diachronic changes in language are unpredictable, with certain terminology evolving from having positive or neutral to offensive connotations (e.g., "mistress") while other terminology evolves from having offensive to positive or neutral connotations (e.g., "queer") [125, 128]. Participants came to a consensus that the best approach to this challenge is to include as much terminology as possible with contextualization, though they also expressed dissatisfaction with how this approach could make descriptions "*clunky-looking*" or "*difficult to read.*"

O6: Barriers posed by institutional and professional structures. While acknowledging the power held as the people describing collections, participants also discussed limitations to their power. Regarding institutional barriers faced by information and heritage professionals when requesting cataloging resources, P1 stated, "*We will always get back to needing an army of catalogers,*" referring to the impossibility of ever hiring enough people to describe a GLAM institution's cultural heritage records to a satisfactory level of detail. Regarding the limited power that information and heritage professionals have over GLAM cataloging standards and expectations for interoperable GLAM systems, participants discussed a tension between standards and the reality of cultural heritage records. P3 stated that social biases are encoded not only in cataloging standards and systems built to use those standards, but also "*the whole profession.*" P1 asked rhetorically, "*Who's on the boards that get to deem that this [e.g., the standards' terminology] is the right language?*" Implicit in participants' discussion was the assumption that social biases are inevitable. Nonetheless, they spoke of ways to use the tools of their profession to contribute to efforts that empower, rather than further oppress, marginalized communities of people. P1 stated, "*We were taught to accept them [the standards] and that's, well, I don't want to. I want to see them as a framework.*"

O7: Desire for more time to document collections. Participants spoke of time constraints that limit the research they can conduct on cultural heritage records to inform their descriptions of the records in catalogs. P3 spoke of the implications of digital technologies, stating, "*For digital archives, at the scale some of them are, the human race will die out before we'd be able to catalog every website.*" P8 spoke of the mismatch between the time needed and the time available for describing cultural heritage records that the participant was currently cataloging: "*[...] then you have to research her relationship, his relationship to get a hint as to what Mary's role was as well, and that takes you into a three year Ph.D.! And you've got six months to*

catalog the collection!" Participants saw an unresolvable tension between making as many records as possible discoverable (*i.e.*, writing descriptions that researchers, students, and the general public can query in online catalogs) and conducting lengthier research into cultural heritage records that could challenge the social biases they may reflect.

6.2 RQ2: ML for Informing Understandings and Management of Bias

Regarding observations relevant to RQ2, participants' comments communicated an interest in integrating ML into existing workflows but not in fully automating description writing or description review processes.

O8: Value of gendered language codes individually vs. in combination for understanding collections' biases. Participants expressed agreement regarding the application of the **Gendered Role** code. P8 communicated an interest in seeing the quantity of different text spans coded with **Gendered Role**, stating that those quantities could support the search for "*bias and gaps*" in a catalog's descriptions. P8 explained that this code would be especially interesting for identifying "*where a 'Mrs.' is, and saying that's potentially a partner that needs further exploration,*" referring to gaps in historical documentation about women's work and contributions [7, 48, 59].

Participants expressed disagreement regarding the application of the **Gendered Pronoun** and **Unknown** codes. P2 and P5 expressed interest in seeing gendered pronouns annotated if there was an assumption being made about a gender group implicitly communicated through the use of the pronoun (*i.e.*, the language that the **Generalization** code is intended to annotate). These participants were not interested in seeing the code applied to every instance of a gendered pronoun across a catalog's descriptions. P7 expressed disagreement with the **Unknown** code's application to one instance of a person's name being based on grammatically or lexically gendered terminology provided only in the description in which that instance of the name appeared, rather than across all descriptions in which that person's name appeared. P10 communicated a similar disagreement, pointing to the third description (Figure 3) in which "Elizabeth II" had been coded as **Unknown**: "*In this case it's a woman in power, but then we're not recognizing that there were some women in power, if everybody's not identified by gender.*" This disagreement relates to participants' sentiments about the value of gendered language as summarized in O1.

Conversely, P1 expressed agreement with the codes' application when considered in combination. P1 stated, "*It's not about just the Gendered Pronoun, it's actually when they come together, they produce something that's really useful.*" This participant felt that the codes as a whole contributed to a better understanding of the gender biases in cultural heritage records and descriptions of them in catalogs, because together the codes provide a large-scale interpretation of a catalog's descriptive language. P1 viewed the ML models and their classification of descriptions with these codes as tools to incorporate into information and heritage professionals' workflows, with the information and heritage professionals determining when and how to apply the tools.

O9: Cautious interest in using ML to support investigations into collections' biases. Participants communicated both skepticism and interest in the ML models' ability to inform their understanding of gender biases in collections and collections' descriptions. P1 proposed, “*I wonder whether actually, if there were collections we would suggest putting through the models to test them, to see if they pick up on things that we know are in those collections, whether that would be something useful.*” Other participants confirmed that they would be interested in reviewing model-made codes on descriptions of cultural heritage collections that they were familiar with to evaluate the ML models' relevance to their understanding of collections' gender biases. That being said, some participants saw the ML models' classifications of potentially gender biased language useful as an initial step for critical cataloging workflows. Regarding applications of the models, P9 said, “*it would maybe enable this kind of revisiting and correcting to be done more quickly and with less labor.*” Critical cataloging could be a use case for human-ML collaboration, with the ML models selecting a subset of a catalog's descriptions for an information or heritage professional to review, rather than the information or heritage professional having to start with the entire catalog.

O10: Models as a tool to integrate into existing workflows to improve awareness of collections' biases. Participants explained that upon being presented with ML models' classifications of gender biased language in descriptions, the next step they would want to take would be to have an information or heritage professional manually review the classified descriptions. P5 viewed the models as, “*a way to flag something that might be problematic. But then you've got to have an archivist or cataloger go back,*” rather than using the models as the final decision-maker on whether additions or changes should be made to a description. Participants also saw the models' classifications as useful for informing guidelines for describing cultural heritage records, particularly for students and volunteers who are writing descriptions. P2 proposed that the models' output could inform a “*how-to guide*” for describing collection material, to which P8 responded, “*I think that would be quite useful to have,*” explaining how such a guide would be helpful for that participant's current work with volunteer catalogers.

O11: Models as a tool for self-reflection on biases encoded in one's own language choices for descriptions. Participants also discussed the value of the models for facilitating their own reflection on how they describe cultural heritage collections. Participants spoke of how concepts of bias evolve and how different stakeholders of a collection may have different opinions about which terminology is correct and which is biased, as further described in O5. Rather than trying to fix or remove bias, P1 proposed that the aim could be awareness, with the ML models helping participants in “*being aware of our practice, or aware of where we've taken old catalogs and perpetuated things.*”⁶

⁶When a GLAM institution acquires a new collection of cultural heritage records, depending on the resources the institution has available, information and heritage professionals may quote from descriptions the collector (*i.e.*, the person or people from whom the institution acquired the collection) already wrote of the cultural heritage records in the institution's catalog, because GLAM institutions' primary aim is to make records discoverable as quickly as possible.

O12: Model outputs as evidence to support resource requests. P1 spoke of how the models could provide evidence of the need for particular description and cataloging projects. P1 discussed the models' classifications as “*an evidence base*” to support requests for hiring interns and full-time information and heritage professionals. This participant saw the models' coding of descriptions as a way to demonstrate that, “*there is a skill to this [description] and these skills of archivists, librarians, and curators are needed in this space.*”

O13: Uncertainty about ML models' capabilities. Upon reviewing the tables of collections with the highest counts of descriptions that were model-coded with **Omission** or **Stereotype**, P1 expressed interest in three of the collections that worksheet 2 listed (Figure 4). Those collections' presence in the tables aligned with P1's knowledge of them as containing records regarding eugenics, misogyny, imperialism, and gender stereotyping. P1 stated, “*I'm absolutely fascinated that Godfrey Thomson⁷ has come up,*” explaining that the high **Stereotype** counts could be due to “*the time period he worked in but also the environment. I think some of his work early on looked at eugenics and that kind of classification of people.*” That being said, P1 expressed uncertainty about whether the models had labeled these collections for the reasons she expected, stating, “*we'd have to investigate further.*”

P7's comments expressed the strongest doubt in the models' capabilities relative to those of the other participants. In regards to the presence of the *Roslin Slide Collection*⁸ in the tables recording collections with high **Omission** and **Stereotype** counts, P7 stated, “*I suspect there is some complicated problem with the data and the collection, and the way the data is being compiled, which is why it's got such high numbers there.*” P1 responded that the presence of the *Roslin* collection in the tables did not surprise her because “*The images [in the collection] were collected during the height of empire and colonialism, and there is language within it that is gendered. There's also how women are represented and that can be in terms of different parts of the world and communities.*”

O14: Little interest in quantitative model evaluation measures. Participants did not ask questions or comment on the second worksheet's charts about the models' performance measured with metrics that inform standard ML evaluations, namely true positives, false positives, and false negatives (Figure 5). Handwritten notes on participants' worksheets suggest that participants were more interested in gaining an understanding of the type of language that the models classified with the taxonomy's codes.

O15: Distinct needs of GLAM employees and visitors. In response to our wrap up question about sharing information on the worksheets that the ML models provided with visitors to the collections (Q3.1), P1 stated, “*I think it'd be really good to ask some of our User Services people who get inquiries in.*” Another participant stated, “*It'd be useful to speak to users [...] because I've got a feeling they come with knowledge that we don't have,*” recalling discussion earlier in the workshop about descriptions of cultural heritage records being works-in-progress intended to help researchers discover potential historical gaps they can investigate further. The results of such an investigation could be brought back to the information

⁷archives.collections.ed.ac.uk/repositories/2/resources/85818

⁸archives.collections.ed.ac.uk/repositories/2/resources/85706

professionals for “enhancing or building upon” a catalog’s existing descriptions.

During the workshop’s pilot, a participant’s comment on the second worksheet’s charts of cross-collection measurements (Figure 5) provide some insight on information a visitor to a catalog might benefit from seeing. The participant stated, “*I just don’t really think about the fact that they [GLAM catalog metadata descriptions] are created in a certain time period by certain people. So I think it adds a layer of something to think about when you’re looking for materials [...] it’s a reminder as a researcher that this is not some sort of neutral, objective tag of what you can actually find.*”

7 Discussion

Predominant approaches to bias in ML aim to measure and remove bias, yet they have not adequately studied how bias may manifest in data. As a result, the inequitable power relationships in society that cause bias in ML systems are rarely identified, let alone challenged, with these approaches. Conversely, the human-centered ML approach we implemented reframes the problem of ML bias, aiming to identify gender bias so it can be communicated. Making bias visible in this way informs and motivates efforts to challenge cisnormative, patriarchal societal structures that create and reinforce gender bias, such as Hessel’s narrative of art history that centers women [59] and the Trans Metadata Collective’s cataloging resource, *Best Practices for Describing Trans and Gender Diverse Resources* [22]. Our work aligns with D’Ignazio’s “restorative/transformative data science,” which leverages data science methods to analyze power, reframe problems, challenge hegemonic worldviews, and produce data [37]. By investigating how bias may manifest in data and the limitations of ML methods for identifying those manifestations, we contribute evidence in support of calls to expand ML bias and fairness approaches [10, 13, 29, 47, 79]. Uniquely, we have offered an approach for determining the feasibility of removing bias or achieving fairness in a given ML use case.

7.1 RQ1: Conceptualizations of Bias

Participants approached our ML models with caution due to concerns about the contextual information that ML models cannot reliably encode. They reject the assumption that computational technologies alone always offer the best solution or approach, a form of bias that Broussard terms “technochauvinism” [18]. Instead, participants viewed the models as valuable for a “human-in-the-loop” system, where “a machine does a lot of the work but meaningful human work and meaningful human interaction are prioritized” (*ibid.*). This definition of *human-in-the-loop* differs from how such systems are described in ML literature, which tends to focus on using human feedback to augment data and manipulate models, integrating human labor into a predominantly ML-driven process [142].

From the 15 observations reported in §6, we identified three characteristics of bias as conceptualized by information and heritage professionals:

- (1) Data bias is contextual, meaning it is a result of complex interactions between cultural, historical, economic, political, and social events, relationships, and practices that are constantly evolving (O3, O4, O5, O6, O15).

- (2) The values or needs of different communities of people may conflict, so the approach to managing data bias that one community prefers may be considered harmful by another community (O1, O5).
- (3) Data bias is inevitable, because the choice of what data to collect and not collect, and how to interpret the collected data, are processes that privilege certain perspectives at the expense of others (O2, O3, O11).

These characteristics motivate information and heritage professionals’ desire to use ML to enhance, rather than fully automate, critical cataloging. Furthermore, they illustrate how information and heritage professionals’ conceptualization of bias, which is based on the external relations of data, differs from predominant conceptualizations of bias among ML researchers and practitioners. Although there are exceptions (e.g., [8]), most ML approaches to bias are based on the internal relations of data, meaning characteristics that are mathematically quantifiable in word embeddings or comparisons of a model’s performance for different demographic groups (§3). Consequently, the ML community continues to deploy models without adequately accounting for the contextual nature of bias, and the models’ widespread adoption harms already-marginalized communities [65]. Though existing research has considered values in ML systems [11, 76], more work is needed investigating value trade-offs, considering how prioritizing one community’s values can harm another community. Additionally, while removing bias and aiming to create fair ML systems is suitable for certain use cases (e.g., [23]), our workshop findings demonstrate that ML approaches to bias must also consider how to manage bias when it cannot or should not be removed.

Notably, participants never spoke of removing bias or neutralizing the language of a description. Instead, participants spoke of how a person’s understanding of what is biased changes over time and varies depending on the relationship between people who create, are represented in, interpret, and use the data. This aligns with feminist understandings of data and knowledge as *situated* [51, 52, 62] and *intersectional* [33, 38], where people’s experiences of privilege and oppression result from the interaction of multiple identity characteristics within particular cultural, historical, economic, political, social, and temporal contexts. For GLAM workflows, choosing what to describe in detail and what to summarize more vaguely inevitably privileges certain points of view over others [41].

Through first-hand experience creating data to describe collections, participants were well aware that managing data biases will always be an imperfect, work-in-progress effort with no universally-applicable solution [17, 25, 136]. Participants acknowledged the power they hold in writing descriptions of heritage collections that people will reference “*as a kind of non-biased factual resource*” (O5), as well as limitations on that power from societal, institutional, and professional structures (O5, O6).

Even when the focus of an area of work is to remove bias, the situated nature of any person or ML model carrying out the work means there will always be some bias, as participants found with the application of the ***Occupation*** code (O2), as Dias Oliva *et al.* have observed in the silencing of the LGBTQ community as a result of efforts to suppress hate speech online with ML [36], and as Parmar *et al.* observed with bias rooted in data annotation instructions

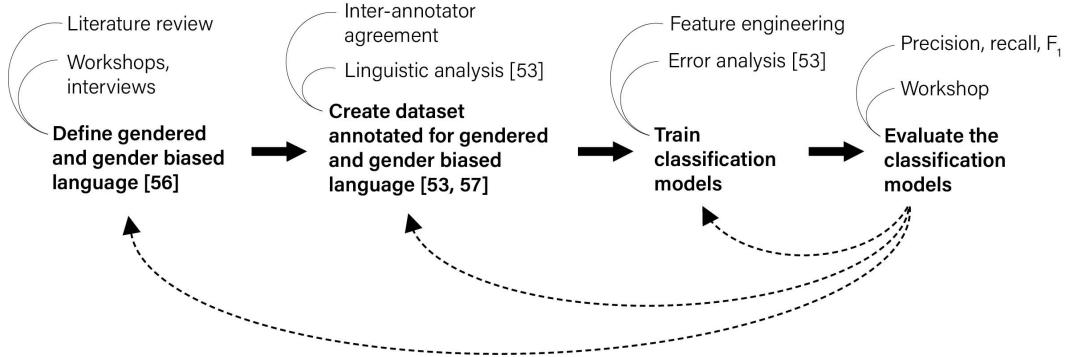


Figure 6: The same flow chart of our four-step research process as Figure 1 with three dotted line arrows added. These new arrows connect the evaluation step to each of the previous steps, indicating the iterations that future work can follow to refine how a model’s task is defined or conceptualized, refine the contents and quality of a dataset, and refine the design of a model. Using traditional ML approaches enables these iterations to be undertaken quickly and at low cost. Once the quantitative and qualitative evaluations of a model are satisfactory, deep learning approaches could be used to create a model on the iteratively refined training dataset.

for crowdworkers [111]. Moreover, in the GLAM domain, removing data bias causes harm by erasing evidence of historical injustice. Records of bias are necessary to learn about equity, fairness, and justice; to hold people and society accountable; and to guide social change [14, 105, 115]. ML systems for Information Retrieval (IR) and summarization have similar purposes as GLAM workflows (*i.e.*, describing and providing access to information), thus our findings are especially relevant to how these ML systems are designed. We encourage IR and summarization research that aims to communicate bias and present more than one perspective to end users, such as news aggregator AllSides’ approach to political bias in news reporting.⁹

7.2 RQ2: ML for Informing Understandings and Management of Bias

From the 15 observations reported in §6, we identified three ways ML can contribute to information and heritage professionals’ understanding and management of bias in GLAM collections and their catalog descriptions:

- (1) Supporting critical cataloging with estimates of the quantities of potentially biased descriptions per collection, and with the identification of specific descriptions most likely to contain gender biased language, informing the prioritization of collections and collection descriptions to review (O8, O9, O10);
- (2) Informing improvements to description practices to avoid perpetuating social biases while staying true to the contents of cultural heritage records (O10, O11); and
- (3) Providing evidence of resource needs for cataloging work, including critical cataloging and writing new descriptions (O7, O12).

Manual description processes cannot keep pace with the rapid rate at which GLAM acquire cultural heritage (O6, O7), so ML models’ ability to process text at large scale can add efficiency to

advance information and heritage professionals’ primary goal of making cultural heritage discoverable. Additionally, if a GLAM institution adopts ML models trained on data that were coded by people outside the institution, the models can draw information and heritage professionals’ attention to biases that they may not have otherwise considered, expanding their understanding of data bias.

7.3 Limitations and Future Work

Our workshop surfaced concerns of participants relevant to future work combining HCI and ML methods to examine data bias. We began the workshop by stating that the aim of identifying gender biases in catalog descriptions was to highlight how structural injustice manifests in language and that the aim was not to place the blame or responsibility for gender biased language on participants. We also reiterated this throughout the workshop. Nonetheless, participants expressed discomfort with evaluating catalog descriptions for bias, describing the workshop discussion as “intimidating” and “self-revealing.” This sentiment may have reduced participants’ willingness to voice their reflections or questions. Following D’Ignazio and Klein’s data feminism principle to “elevate emotion and embodiment” [38], we encourage future work to investigate how to apply human-centered research methods to data bias evaluations that engage domain experts in an empowering manner and minimizes their feelings of discomfort. Aligning with Young’s concepts of structural injustice and social connection model of responsibility [146], we emphasize that while all those working with data have a responsibility to try to counteract structural injustices reflected and perpetuated in data, they are not solely responsible for, nor are they to blame for, biases in those data or any models trained on them.

Future work will include further error analysis and experiments to improve the text classifiers, and evaluations of their performance on additional data. Given the higher precision than recall scores for the Omission and Stereotype Classifiers, error analysis will be conducted to investigate whether the difficult-to-classify instances

⁹allsides.com

represent one or several manifestations of gender biased language. Further feature engineering experiments (e.g., including part of speech tags as features) will be conducted with the aim of improving the classifiers' performance, and to further investigate the extent to which lexically and grammatically gendered language correlates with gender biased language. Running and evaluating the classifiers on datasets in similar and different cultural contexts will provide insight on whether linguistic patterns in bias manifestation hold across different text corpora. To investigate bias transfer and develop approaches to distinguishing biases from models' pre-training and fine-tuning datasets, future work could include experiments with pre-trained models, comparing the gender biases they classify with those of this paper's models. When creating ML and AI systems, we encourage researchers and practitioners to consider using traditional ML and our mixed-methods evaluation approach to iterate quickly and at low cost, working to more thoroughly minimize the risk of harm from unknown biases before investing in the resources needed to train deep learning models. As Figure 6 illustrates, the quantitative and qualitative evaluation of models can inform repeated refinements to the conceptualization of a task or problem, the contents and quality of training data, and model design.

8 Conclusion

While creating unbiased and fair ML systems is a worthy goal, in practice, the contextual, dynamic, and inevitable nature of bias means supposedly fair ML systems may uphold existing, inequitable power relationships in society, risking the perpetuation and amplification of social biases. ML approaches to bias and fairness must expand to consider when the best aim is to manage, rather than eliminate, bias. Through the creation of ML models designed to identify gender biased language, and a human-centered evaluation of the models and their underlying data, we have (a) demonstrated the limitations of predominant ML approaches to bias and fairness, and (b) provided an approach to investigating the feasibility of eliminating bias from an ML system. We encourage ML researchers and practitioners to use our mixed-methods approach from the start of an ML system's design, so the training dataset and model can be iteratively refined quickly and at low cost until the social biases and their consequences for the ML system's use case are thoroughly understood.

Acknowledgments

Thank you to the archivists, librarians, and curators on the University of Edinburgh's Heritage Collections teams, for their contributions to this research, and to the reviewers, whose thoughtful feedback helped us strengthen this paper. We also extend our gratitude to the organizations who provided grants to support our research: the University of Edinburgh's Edinburgh Futures Institute and Informatics Graduate School, and the UKRI EPSRC.

References

- [1] Melissa Adler. 2017. *Cruising the Library: Perversities in the Organization of Knowledge*. Fordham University Press, New York. <https://doi.org/10.2307/j.ctt1xhr79m>
- [2] Christine M. Angel and Caroline Fuchs (Eds.). 2018. *Organization, Representation and Description through the Digital Age: Information in Libraries, Archives and Museums*. Walter de Gruyter GmbH, Boston. <https://doi.org/10.1515/9783110337419>
- [3] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. <https://doi.org/10.15779/Z38BG31>
- [4] Valeria Basile, Federico Cabitzia, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 6 (June 2021), 6860–6868. <https://doi.org/10.1609/aaai.v37i6.25840>
- [5] Christine Basta, Marta R Costa-jussà, and Noe Casas. 2020. Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings. *Neural Computing & Applications* 33, 8 (2020), 3371–3384.
- [6] Eric PS Baumer. 2017. Toward Human-Centered Algorithm Design. *Big Data & Society* 4, 2 (2017), 1–12. <https://doi.org/10.1177/2053951717118854>
- [7] Mary Beard. 2017. *Women & Power: A Manifesto*. Profile Books Ltd, London.
- [8] Kaspar Beelen, Jon Lawrence, Daniel C. S. Wilson, and David Beavan. 2022. Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan. *Digital Scholarship in the Humanities* 38, 1 (July 2022), 1–22. <https://doi.org/10.1093/llc/fqac037>
- [9] Dorothy Berry. 2020. Conscious Editing: Enhancing Diversity and Discovery. <https://youtu.be/XGCTtDgNty4>
- [10] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *SCRIPT-ed* 17, 2 (August 2020), 389–409. <https://doi.org/10.2966/scrip.170220.389>
- [11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, KR) (FAccT '22). ACM, New York, NY, US, 173–184. <https://doi.org/10.1145/3531146.3533083>
- [12] Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa, HI, US, 1536–1546. <https://doi.org/10.1109/WACV48630.2021.00158>
- [13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 5454–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- [14] Francis X. Blouin, Jr. and William G. Rosenberg. 2011. *Processing the Past: Contesting Authorities in History and the Archives* (online ed.). Oxford University Press, Oxford, Chapter 8, 140–160. <https://doi.org/10.1093/acprof:oso/9780199740543.001.0001>
- [15] Cat Bohannon. 2023. *Eve: How the Female Body Drove 200 Million Years of Human Evolution* (first ed.). Alfred A. Knopf, New York.
- [16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (December 2017), 135–146. https://doi.org/10.1162/tacl_a_00051
- [17] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, Cambridge.
- [18] Meredith Broussard. 2023. *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. The MIT Press, Cambridge.
- [19] Mary Bucholtz. 1999. Gender. *Journal of Linguistic Anthropology* 9, 1–2 (1999), 80–83.
- [20] Mary Bucholtz. 2003. Theories of Discourse as Theories of Gender: Discourse Analysis in Language and Gender Studies. In *The Handbook of Language and Gender*. Blackwell Publishing Ltd, Oxford, 43–68. <https://doi.org/10.1002/9780470756942.ch2>
- [21] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, Online, 1–15. <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
- [22] Jasmine Burns, Michelle Cronquist, Jackson Huang, Devon Murphy, K.J. Rawson, Beck Schaefer, Jamie Simons, Brian M. Watson, Adrian Williams, and The Trans Metadata Collective. 2022. Metadata Best Practices for Trans and Gender Diverse Resources. <https://doi.org/10.5281/zendono.6829167>
- [23] Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (online). ACL, online, 4568–4595. <https://doi.org/10.18653/v1/2020.acl-main.418>
- [24] Michelle Caswell. 2022. Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal. In *Radical Empathy in Archival Practice* (special issue on feminist ethics ed.), Vol. 3. Journal of Critical Library and Information Studies, online, 36 pages. Issue 1. <https://journals.litwinbooks.com/index.php/jclis/article/view/113/67>
- [25] Michelle Caswell and Marika Cifor. 2016. From Human Rights to Feminist Ethics: Radical Empathy in the Archives. *Archivaria* 81 (May 2016), 23–43. <https://archivaria.ca/index.php/archivaria/article/view/13557>
- [26] Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models. In

- Proceedings of the 14th International Conference on Natural Language Generation.* ACL, Aberdeen, Scotland, GB, 55–63. <https://aclanthology.org/2021.inlg-1.7>
- [27] Catherine Nicole Coleman. 2020. Managing Bias When Library Collections Become Data. *International Journal of Librarianship* 5, 1 (July 2020), 8–19. <https://doi.org/10.23974/ijol.2020.vol5.1.162>
- [28] Terry Cook. 2011. ‘We Are What We Keep; We Keep What We Are’: Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists* 32, 2 (2011), 173–189. <https://doi.org/10.1080/00379816.2011.619688>
- [29] Sam Corbett-Davies, Johann D. Gaeberl, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2024. The Measure and Mismeasure of Fairness. *The Journal of Machine Learning Research* 24, 1, Article 312 (March 2024), 116 pages.
- [30] Ryan Cordell. 2020. Machine Learning + Libraries: A Report on the State of the Field. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig>
- [31] Kate Crawford. 2021. *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven.
- [32] Kate Crawford and Trevor Paglen. 2019. *Excavating AI: The Politics of Training Sets for Machine Learning*. The AI Now Institute, NYU, New York, NY, US. <https://excavating.ai>
- [33] Kimberlé Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* 1989, 1 (1989), 139–167. <https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8>
- [34] Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics and Violence Against Women of Color. *STANFORD LAW REVIEW* 43, 6 (1991), 60. <https://doi.org/10.2307/1229039>
- [35] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. ACM, virtual event (US), 1591–1602. <https://doi.org/10.1145/3461778.3462131>
- [36] Thiago Dias Oliva, Dennys Marcelo Antonioli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (2021), 700–732. <https://doi.org/10.1007/s12119-020-09790-w>
- [37] Catherine D’Ignazio. 2024. *Counting Feminicide: Data Feminism in Action*. The MIT Press, Cambridge. <https://doi.org/10.7551/mitpress/14671.001.0001>
- [38] Catherine D’Ignazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press, Cambridge. <https://mitpressonpubpub.mitpress.mit.edu/data-feminism>
- [39] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Online, 314–331. <https://www.aclweb.org/anthology/2020.emnlp-main.23>
- [40] Carl DiSalvo. 2022. *Design as Democratic Inquiry: Putting Experimental Civics into Practice*. The MIT Press, Cambridge. <https://doi.org/10.7551/mitpress/13372.001.0001>
- [41] Wendy M. Duff and Verne Harris. 2002. Stories and Names: Archival Description as Narrating Records and Constructing Meanings. *Archival Science* 2, 3–4 (September 2002), 263–285. <https://doi.org/10.1007/BF02435625>
- [42] Gordon Dunsire. 2018. Ethical Issues in Catalogue Content Standards. *Catalogue and Index June 2018* (2018), 11–15. Issue 191.
- [43] Norman Fairclough. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Routledge, London.
- [44] Jonathan Turner. 2007. Dewey Deracialized: A Critical Race-Theoretic Perspective. *Knowledge Organization* 34, 3 (2007), 144–168. <https://doi.org/10.5771/0943-7444-2007-3-144>
- [45] Barney G. Glaser and Anselm L. Strauss. 1980. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, New York.
- [46] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, Online, 1926–1940. <https://doi.org/10.18653/v1/2021.acl-long.150>
- [47] Samuel Goree and David Crandall. 2023. Situated Cameras, Situated Knowledges: Towards an Ego-centric Epistemology for Computer Vision. In *3rd International Ego4D Workshop @ CVPR 2023*, Vol. abs/2307.00064. ArXiv, Online, 4 pages. <https://doi.org/10.48550/arxiv.2307.00064>
- [48] David Graeber and D. Wengrow. 2021. *The Dawn of Everything: A New History of Humanity*. Penguin Books Ltd., London, England, GB.
- [49] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. (2023). <https://doi.org/10.2139/ssrn.4543807> [accepted to appear in Nature].
- [50] Mark A. Greene and Dennis Meissner. 2005. More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist* 68, Fall/Winter 2005 (2005), 208–263. <https://www2.archivists.org/sites/all/files/MPLP-AmericanArchivist-2005.pdf>
- [51] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575 pages. <https://doi.org/10.2307/3178066>
- [52] Sandra Harding. 1995. “Strong Objectivity”: A Response to the New Objectivity Question. *Synthese* 104, 3 (September 1995), 331 pages. <https://doi.org/10.1007/BF01064504>
- [53] Lucy Havens. 2024. *Recalibrating Machine Learning for Social Biases: Demonstrating a New Methodology through a Case Study Classifying Gender Biases in Archival Documentation*. Ph.D. Dissertation. University of Edinburgh.
- [54] Lucy Havens. 2025. Data Bias and the Natural Language Processing of Metadata. In *Library Catalogues as Data: Research, Practice, and Usage*. Facet Publishing, London, chapter 4.
- [55] Lucy Havens, Melissa Terras, and Beatrice Alex. 2025. Confronting Gender Bias in Heritage Catalogues: A Natural Language Processing Approach to Revisiting Descriptive Metadata. In *The Routledge Handbook of Heritage and Gender (Routledge Handbooks on Museums, Galleries and Heritage)*. Routledge, GB, 335–350. <https://doi.org/10.4324/9781003258193-31>
- [56] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated Data: Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. ACL, Barcelona, Spain (Online), 107–124. <https://aclanthology.org/2020.gebnlp-1.10>
- [57] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. ACL, Seattle, WA, US, 30–57. <https://doi.org/10.18653/v1/2022.gebnlp-1.4>
- [58] Denise Hearn. 2024. *Harms from Concentrated Industries: A Primer*. Technical Report. Columbia Center on Sustainable Development.
- [59] Katy Hessel. 2023. *The Story of Art Without Men* (first American ed.). W.W. Norton & Company, Inc., New York, NY, US.
- [60] Mar Hicks. 2018. *Introduction: Britain’s Computer “Revolution”*. The MIT Press, Cambridge, 1–17.
- [61] Mar Hicks. 2021. *Sexism is a Feature, Not a Bug*. The MIT Press, Cambridge, Chapter 6, 135–158.
- [62] Patricia Hill Collins. 2000. *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*. Routledge, New York.
- [63] Uta Hinrichs, Stefania Forlini, and Bridget Moynihan. 2019. In Defense of Sandcastles: Research Thinking Through Visualization in Digital Humanities. *Digital Scholarship in the Humanities* 34, Supplement 1 (Dec. 2019), i80–i99. <https://doi.org/10.1093/llc/fqy051>
- [64] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society* 22, 7 (June 2019), 900–915. <https://doi.org/10.1080/1369118X.2019.1573912>
- [65] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Shareen King. 2024. AI Generates Covertly Racist Decisions About People Based on Their Dialect. *Nature* 633, 8028 (September 2024), 147–154. <https://doi.org/10.1038/s41586-024-07856-5>
- [66] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*. ACM, Glasgow, Scotland, GB, 1–16. <https://doi.org/10.1145/329065.3300830>
- [67] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal of Responsible Computing* 1, 2 (2024), 52 pages. <https://doi.org/10.1145/3631326>
- [68] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design For and With Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Ft. Lauderdale, FL, US) (CHI ’03)*. ACM, New York, NY, US, 17–24. <https://doi.org/10.1145/642611.642616>
- [69] ICA. 2011. ISAD(G): General International Standard Archival Description - Second Edition. <https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition>
- [70] Lise Jaillant (Ed.). 2022. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld University Press, Bielefeld, DE. <https://doi.org/10.14361/9783839455845>
- [71] Sophie Jentszsch and Cigdem Turan. 2022. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. ACL, Seattle, WA, US, 184–199. <https://doi.org/10.18653/v1/2022.gebnlp-1.20>
- [72] Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Investigating the Capabilities and Limitations of ML for Identifying Bias*

- Human Language Technologies.* ACL, Online, 3770–3783. <https://doi.org/10.18653/v1/2021.nacl-main.296>
- [73] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency* (Barcelona, Spain) (FAT*’20). ACM, New York, NY, US, 306–316. <https://doi.org/10.1145/3351095.3372829>
- [74] Samia Kabin Lixiang Li, and Tianyi Zhang. 2024. STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, US) (CHI ’24). ACM, New York, NY, US, Article 293, 20 pages. <https://doi.org/10.1145/3613904.3642111>
- [75] Atoosa Kasirzadeh. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford, England, GB, 349–356. <https://doi.org/10.1145/3514094.3534188>
- [76] Atoosa Kasirzadeh and Jason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology* 36, 27 (April 2023), 1–24. <https://doi.org/10.1007/s13347-023-00606-x>
- [77] Lena Kástner and Barnaby Crook. 2024. Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science* 14, 4 (2024), 52. <https://doi.org/10.1007/s13194-024-00614-4>
- [78] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 88 (November 2018), 22 pages. <https://doi.org/10.1145/3274357>
- [79] Lauren Klein and Catherine D’Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro, Brazil, 100–112. <https://doi.org/10.1145/3630106.3658543>
- [80] Mikhail Korobov. 2015. sklearn-crfsuite. <https://sklearn-crfsuite.readthedocs.io/en/latest> [online; accessed June 6, 2023].
- [81] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatiariaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18*. ACM Press, Lyon, France, 853–862. <https://doi.org/10.1145/3178876.3186133>
- [82] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology*. SAGE, Los Angeles.
- [83] Kaysi Eastlick Kushner and Raymond Morrow. 2003. Grounded Theory, Feminist Theory, Critical Theory: Toward Theoretical Triangulation. *Advances in Nursing Science* 26 (2003), 30–43. <https://doi.org/10.1097/00012272-200301000-00006>
- [84] Faisal Ladzhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 3206–3219. <https://aclanthology.org/2023.eacl-main.234>
- [85] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, US) (FAccT ’23). ACM, New York, NY, US, 1369–1385. <https://doi.org/10.1145/3593013.3594087>
- [86] Anna Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 85–91. <https://doi.org/10.18653/v1/S19-1010>
- [87] LC Labs. n.d. Exploring Computational Description. <https://labs.loc.gov/work-experiments/ECD/?loclr=blogsig> [Accessed November 20, 2024].
- [88] Patricia Leavy. 2017. *Qualitative Research Design*. Guilford Publications, New York, 131–163. <https://ebookcentral.proquest.com/lib/ed/reader.action?docID=4832778&ppg=244>
- [89] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu, HI, US, 1–15. <https://doi.org/10.1145/3313831.3376590>
- [90] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790 [cs.AI] <https://arxiv.org/abs/2110.10790>
- [91] Tom Liberum, Senthooran Rajamanoharan, Conny Arthur, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. *Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2*. Technical report, Google DeepMind. <https://storage.googleapis.com/gemma-scope/gemma-scope-report.pdf>
- [92] Library of Congress. n.d. Fascinating Facts | About the Library | Library of Congress. <https://www.loc.gov/about/fascinating-facts> [Accessed December 1, 2024].
- [93] Library of Congress. n.d.. General Information | About the Library | Library of Congress. <https://www.loc.gov/about/general-information/> [Accessed December 1, 2024].
- [94] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwisch, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama, Japan, 1–14. <https://doi.org/10.1145/3411764.3445488>
- [95] Edward Lopez and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, PA) (ETMTNLP ’02). ACL, US, 63–70. <https://doi.org/10.3115/1118108.1118117>
- [96] Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. 2020. Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project. https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf
- [97] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*. ACL, Virtual, 48–55. <https://doi.org/10.18653/v1/2021.nuse-1.5>
- [98] Richard Marciano. 2022. *Afterword: Towards a New Discipline of Computational Archival Science (CAS)*. Bielefeld University Press, Bielefeld, 205–218.
- [99] Nina Markl and Catherine Lai. 2021. Context-Sensitive Evaluation of Automatic Speech Recognition: Considering User Experience & Language Variation. In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. ACL, Online, 34–40. <https://aclanthology.org/2021.hcincpl-1.6>
- [100] Melissa McCradden, Mjaye Mazwi, Shalmali Joshi, and James A. Anderson. 2020. *When Your Only Tool is a Hammer: Ethical Limitations of Algorithmic Fairness Solutions in Healthcare Machine Learning*. ACM, New York, NY, US, 109. <https://doi.org/10.1145/3375627.3375824>
- [101] Ninhare Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6, Article 115 (July 2021), 35 pages. <https://doi.org/10.1145/3457607>
- [102] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsudeen Hassan Muhammed, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Oreaoghemah Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahe, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyejulu, Chris Emuze, Bonaventure Dosso, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Oktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. ACL, Online, 2144–2160. <https://www.aclweb.org/anthology/2020.findings-emnlp.195>
- [103] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.
- [104] Mahsan Nourani, Amal Hashky, and Eric D. Ragan. 2024. User Profiling in Human-AI Design: An Empirical Case Study of Anchoring Bias, Individual Differences, and AI Attitudes. *Proceedings of the Twelfth AAAI Conference on Human Computation and Crowdsourcing (HCMP 2024)* 12, 1 (October 2024), 137–146. <https://doi.org/10.1609/hcemp.v12i1.31608>
- [105] Society of American Archivists. 2020. SAA Core Values Statement and Code of Ethics. <https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics> [online; accessed September 10, 2024].
- [106] Hope A. Olson. 2001. The Power to Name: Representation in Library Catalogs. *SigNS: Journal of Women in Culture and Society* 26, 3 (April 2001), 639–668. <https://doi.org/10.1086/495624>
- [107] OpenAI. 2024. ChatGPT. <https://chat.openai.com> [online; accessed January 19, 2024].
- [108] Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. ACL, Seattle, WA, 151–167. <https://doi.org/10.18653/v1/2022.gebnlp-1.17>
- [109] Alexandra Ortola-Baird and Julianne Nyhan. 2022. Encoding the Haunting of an Object Catalogue: On the Potential of Digital Technologies to Perpetuate or Subvert the Silence and Bias of the Early-Modern Archive. *Digital Scholarship in the Humanities* 37, 3 (August 2022), 844–867. <https://doi.org/10.1093/llc/fqab065>
- [110] Rock Yuren Pang, Jack Cenatempo, Franklin Graham, Bridgette Kuehn, Maddy Whisenant, Portia Botchway, Katie Stone Perez, and Allison Koenecke. 2023. Auditing Cross-Cultural Consistency of Human-Annotated Labels for Recommendation Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago, IL, US, 1531–1552. <https://doi.org/10.1145/3593013.3594098>

- [111] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 1779–1789. <https://doi.org/10.18653/v1/2023.eacl-main.130>
- [112] Amandalyne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2, 11 (November 2021), 14 pages. <https://doi.org/10.1016/j.patter.2021.100336>
- [113] Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, London.
- [114] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 55, 3, Article 51 (February 2022), 44 pages. <https://doi.org/10.1145/3494672>
- [115] David Pilgrim. 2024. Why I Collect Racist Objects - Jim Crow Museum. <https://jimcrowmuseum.ferris.edu/collect.htm> [online; accessed February 11, 2024].
- [116] Miriam Posner. 2016. *What's Next? The Radical, Unrealized Potential of Digital Humanities*. University of Minnesota Press, Minneapolis, 32–41. <https://doi.org/10.5749/j.ctt1cm6thb.6>
- [117] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalyne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1. Curran, Online, 17 pages. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf
- [118] Jane Ritchie and Jane Lewis. 2003. *In-Depth Interviews*. SAGE Publications Ltd, London, 138–168.
- [119] Colin Robson and Kieran McCartan. 2016. *The Analysis and Interpretation of Qualitative Data*. Wiley-Blackwell, Chichester, 459–486.
- [120] Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case Study: Predictive Fairness to Reduce Misdeemeanor Recidivism Through Social Service Interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 142–153. <https://doi.org/10.1145/3351095.3372863>
- [121] Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting Unintended Social Bias in Toxic Language Datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. ACL, Abu Dhabi, United Arab Emirates (Hybrid), 132–143. <https://aclanthology.org/2022.conll-1.10>
- [122] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, CA) (FAccT '21). ACM, New York, NY, US, 315–328. <https://doi.org/10.1145/3442188.3445896>
- [123] Michele Samorani, Shannon L. Harris, Linda Goler Blount, Haibing Lu, and Michael A. Santoro. 2022. Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. *Manufacturing & Service Operations Management* 24, 6 (2022), 2825–2842. <https://doi.org/10.1287/msom.2021.0999>
- [124] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. ACL, Valencia, ES, 1–10. <https://doi.org/10.18653/v1/W17-1101>
- [125] Muriel R. Schulz. 2000. The Semantic Derogation of Women. In *The Routledge Language and Cultural Theory Reader*. Routledge, London.
- [126] Joan M. Schwartz and Terry Cook. 2002. Archives, Records, and Power: The Making of Modern Memory. *Archival Science* 2, 1–2 (2002), 1–19. <https://doi.org/10.1007/BF02435628>
- [127] Ben Shneiderman. 2022. *Human-Centered AI*. Oxford University Press, New York.
- [128] Norena Shopland. 2020. *A Practical Guide to Searching LGBTQIA Historical Records*. Taylor & Francis Group, Milton. <https://doi.org/10.4324/9781003006787>
- [129] Laurajane Smith. 2006. *Uses of Heritage*. Routledge, London.
- [130] Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19, 3 (1989), 387–420. <https://doi.org/10.1177/030631289019003001>
- [131] Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, Dublin, Ireland, 3524–3542. <https://doi.org/10.18653/v1/2022.acl-long.247>
- [132] Ann Laura Stoler. 2002. Colonial archives and the arts of governance. *Archival Science* 2, 1–2 (2002), 87–109. <https://doi.org/10.1007/BF02435632>
- [133] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, Florence, Italy, 1630–1640. <https://doi.org/10.18653/v1/P19-1159>
- [134] Marja Liisa Swantz. 2008. *Participatory Action Research as Practice*. SAGE Publications Ltd, London, Chapter 2, 31–48. <https://doi.org/10.4135/9781848607934.n8>
- [135] Mohammad Tahaei, Marios Constantinides, Daniela Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q. Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, Ewa Luger, Michael Madaio, Ilana Golbin Blumenfeld, Maria De-Arteaga, Jessica Vitak, and Alexandra Olteanu. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI EA '23). ACM, New York, NY, US, Article 515, 4 pages. <https://doi.org/10.1145/3544549.3583178>
- [136] Jessica Tai. 2021. Cultural Humility as a Framework for Anti-Oppressive Archival Description. *Journal of Critical Library and Information Studies* 3, 2 (October 2021), 23 pages. <https://doi.org/10.24242/jcls.v3i2.120>
- [137] Melissa Terras, James Baker, James Hetherington, David Beavan, Martin Zalta Austwick, Anne Welsh, Helen O'Neill, Will Finley, Oliver Duke-Williams, and Adam Farquhar. 2018. Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High-Performance Computing, and Transforming Access to British Library Digital Collections. *Digital Scholarship in the Humanities* 33, 2 (June 2018), 456–466. <https://doi.org/10.1093/llc/fqx020>
- [138] Theo Thomassen. 2002. A First Introduction to Archival Science. *Archival Science* 1 (2002), 373–385.
- [139] Nanna Bonde Thystrup, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel (Eds.). 2021. *Uncertain Archives: Critical Keywords for Big Data*. The MIT Press, Cambridge.
- [140] Ciaran B. Trace. 2022. Archival Infrastructure and the Information Backlog. *Archival Science* 22, 1 (2022), 75–93. <https://doi.org/10.1007/s10502-021-09368-x>
- [141] Kiri Wagstaff. 2012. Machine Learning that Matters. In *Proceedings of the 29th International Conference on Machine Learning* (Edinburgh, Scotland, UK) (ICML '12). icml.cc / Omnipress, Madison, WI, US, 6 pages. <https://icml.cc/2012/papers/298.pdf>
- [142] Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting Humans in the Natural Language Processing Loop: A Survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang (Eds.). Association for Computational Linguistics, Online, 47–52. <https://aclanthology.org/2021.hcinlp-1.8>
- [143] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics* 6 (2018), 605–617. https://doi.org/10.1162/tacl_a_00240
- [144] Anne Welsh. 2016. The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly* 54, 5–6 (August 2016), 317–337. <https://doi.org/10.1080/01639374.2016.1188433>
- [145] Anne Welsh and Sue Batley. 2009. *Practical Cataloguing: AACR, RDA and MARC21*. Facet Publishing, London.
- [146] Iris Marion Young. 2011. *Responsibility for Justice*. Oxford University Press, Cary. <https://doi.org/10.1093/acprof:oso/9780195392388.001.0001>
- [147] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. <https://doi.org/10.1145/3544548.3580945>
- [148] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordóñez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. ACL, New Orleans, LA, US, 15–20. <https://doi.org/10.18653/v1/N18-2003>

A Machine Learning

Here we detail our machine learning approach for reproducibility.

A.1 Data Preprocessing

To preprocess the data we selected for training ML models (*i.e.*, to make the descriptions machine readable), we removed the metadata field names (*e.g.*, "Title," "Scope and Contents") from the descriptions and converted the text to lowercase; we kept punctuation and stop words in the dataset. We tokenized the input data for the token and sequence classification models with NLTK's Punkt tokenizer [95]. For the sequence classifier, the **Feminine**, **Masculine**,

Non-binary, **Unknown**, and **Occupation** codes were split into beginning and ending tags for the coded text spans' tokens (e.g., the text span "Queen Elizabeth" coded as "Feminine" would have a "B-Feminine" tag on "Queen" and "I-Feminine" tag on "Elizabeth"), as is standard in data preprocessing for named entity recognition. For the sequence classifier, we grouped the tokens by sentence, added START and END booleans to the beginning and end of each sentence, and included a bias value of 1.0 as a feature, as recommended in scikit-learn crf-suite's documentation [80]. For the token and sequence classifiers, we represented the tokens within the descriptions as custom fastText word embeddings [16] of 100 dimensions.¹⁰ For the document classifier, we represented complete descriptions as Term Frequency-Inverse Document Frequency (TF-IDF) matrices.¹¹

A.2 Model Design, Training, and Testing

We created three types of models: token, sequence, and document classification models, all trained on the training dataset described in §4.1. The token classification models assigned individual tokens (words and subwords) codes for the gender association of individual tokens, specifically **Gendered Pronoun**, **Gendered Role**, and **Generalization**. The sequence classification models assigned a running sequence of tokens (*i.e.*, a phrase or other subset of a description) codes for the grammatical or lexical gender associated with a person's name or job title, specifically **Feminine**, **Masculine**, **Non-binary**, **Unknown**, or **Occupation**. We used sequence instead of token classification for these labels due to the longer length of names and job titles' text spans, which ranged from one to ten tokens. The document classification models assigned entire descriptions codes for gender biased language, specifically **Omission** and **Stereotype**. The token and document models were trained in a multilabel task, meaning a model could assign zero, one, or many labels to a token or document. The sequence models were trained in a multiclass task, meaning a model could assign at most one label to a token, because a token cannot be both a job title and a person, nor should a single instance of a person's name be referred to with terms having multiple grammatical or lexical genders (this does allow for different instances of a person's name to have different grammatical and lexical gender associations, however).

Our code was written in Python.¹² We used scikit-learn and scikit-learn-based libraries for all the classification models reported in this paper. The multilabel token classification model is a Classifier Chain¹³ of Random Forests,¹⁴ with the parameter "random_state" to "22" and the remaining parameters set to scikit-learn's defaults. We refer to this model as the Linguistic Classifier (LC), because it classifies with codes from the taxonomy's Linguistic category: **Gendered Pronoun**, **Gendered Role**, and **Generalization**. The multiclass sequence classification models are Conditional Random Fields (CRF) models using the AROW algorithm with the parameters "variance" set to "1," "max_iterations" set to "100," and

"all_possible_transitions" set to "True." We refer to these model as the Person Name and Occupation Classifiers (PNOCs) because it classifies with codes from the taxonomy's Person Name category (**Feminine**, **Masculine**, and **Unknown**) as well as the **Occupation** code from the taxonomy's Contextual category. The multilabel document classification models use Support Vector Machines with Stochastic Gradient Descent¹⁵ in a one-vs.-rest setup with scikit-learn's default parameters. We refer to these models as the Omission and Stereotype Classifiers (OSCs) because they classify descriptions **Omission** and **Stereotype** codes from the with the taxonomy's Contextual category. To arrive at these model setups, we experimented with several algorithms and parameter settings for token, sequence, and document classification, choosing the algorithms and parameters that resulted in the highest F₁ scores (for all experiment detail, see [53]).

To train and test each classification model, we took a modified approach to five-fold cross-validation to generate predictions (model-made codes) for the entire dataset of descriptions. First, we split the dataset into five subsets, or folds, where each fold accounted for a unique 20% selection of the dataset. Then, we iteratively selected four folds (80% of the dataset) for training and used the remaining fold (20% of the dataset) for testing. This approach trained five instances of every model and enabled the generation of predictions for the entire dataset, 20% at a time. We evaluated the models using precision, recall, and F₁ scores, optimizing for F₁ score. These are the same scores calculated to measure the human coders' Inter-Annotator Agreement (IAA) for our training dataset's source data [57], enabling us to compare the automated and manual coding processes. Our code is available on GitHub.¹⁶ We report on the performance of the models using precision, recall and F₁ scores, optimizing for F₁ score.

A.3 Model Performance

We evaluated each individual text classification model across the three model cascades and report the performance scores per label in tables 3, 4, and 5. Table 2 reports the overall performance scores for the OSCs, the last model in each cascade. Given space constraints and the focus of our work on leveraging the models to investigate information professionals' conceptualization of gender bias, we give only a brief discussion of the implications of the models' performance scores here (see [53] for further detail).

The LC yielded the highest performance scores across all models (Table 4). This result is not surprising given the high agreement between human coders in the training dataset for two of the codes that the LC classifies tokens with: **Gendered Pronoun** (F₁=0.957) and **Gendered Role** (F₁=0.780). Performance scores for **Generalization**, however, are lowest across all models' per code scores. Though the model's precision score is reasonably high, at 0.646 (64.6% of all the model's classified instances were meant to be classified), its recall score is very low, at 0.232 (23.2% of all instances that were meant to be classified were, in fact, classified by the model), leading to a low (below 50%) F₁ score of 0.341. This pattern follows the human coders' agreement with the training dataset, which also had high

¹⁰radimrehurek.com/gensim/models/fasttext.html

¹¹scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

¹²www.python.org

¹³scikit.ml/api/skmultilearn.problem_transform.cc.html

¹⁴scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

¹⁵scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

¹⁶github.com/thegeose20/gender-bias

precision and low recall scores. These results indicate that there are different language patterns that indicate an instance of gender generalization and a portion of these patterns were not learned or not learned well by the model. This reflects the confusion that human coders reported during their discussion of how to apply the **Generalization** code throughout the human coding process.

The PNOCs yielded low per code scores relative to the LC and OSCs, with F_1 scores ranging from 0.406 to 0.641 (Table 5). Despite the greater presence of men in the archival catalog and the collections it documents, the **Masculine** code had the lowest score amongst all PNOC codes. Error analysis on the PNOC and human coders' classified data found that **Masculine** and **Unknown** labels were frequently confused. This indicates a high difficulty of the human coding task: though the taxonomy instructed human coders to label names without a grammatically or lexically gendered referent (e.g., "he" or "Sir") as **Unknown**, many person names without such a referent were coded as **Masculine**. This could be due to previous descriptions a human coder had read about the same person that had a grammatically or lexically masculine referent or due to stereotypical inferences made about a person based on their name or other information provided about them in a description, such as their job title. The human coders did not find any instances of grammatically or lexically non-binary referents, resulting in no **Non-binary** codes in the training dataset for models to learn from; this is thus a gap in our model that can be addressed in future work by expanding our dataset. Note that there may still be people in our dataset who identified as non-binary or another gender diverse identity. Due to the marginalization and oppression of people of

non-binary and gender diverse identities, people may have chosen not to document or to destroy records that documented their gender identity [128].

Among the three cascades' OSCs, the OSC for cascades 2 and 3 yielded the highest per label scores. Cascade 3's OSC had the highest F_1 score for **Omission**, at 0.682, which was slightly lower than the human coders' agreement with the aggregated (training) dataset at ($F_1=0.767$) and much higher than the human coders' IAA ($F_1=0.439$). Cascade 2's OSC had the highest **Stereotype** F_1 score at 0.841, which is higher than the human coders' IAA ($F_1=0.440$) and human coders' agreement with the training dataset ($F_1=0.823$). Across all labels, Cascade 3's OSC yielded the highest F_1 scores (macro $F_1=0.761$, micro $F_1=0.730$), though Cascade 2's OSC was a close second (macro $F_1=0.760$, micro $F_1=0.727$). Given the better performance scores of the LC relative to the PNOC, we posit that Cascade 2's OSC is more reliable, because Cascade 2 included the LC's codes as features while Cascade 3 included the PNOC's codes as features. Surprisingly, the inclusion of both the LC and PNOC's codes as features for Cascade 1 yielded the lowest OSC performance, with this OSC yielding macro and micro F_1 scores within 0.001 of the baseline OSC. We also note that the inclusion of taxonomy codes reduced the precision scores (the baseline OSC's precision scores were highest) but improved the recall scores. There was a greater improvement to recall than reduction in precision, however, leading to higher overall OSC performance as measured with macro and micro F_1 scores.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 4: Comparing human coders and text classification model's performance for *Gendered Pronoun*, *Gendered Role*, and *Generalization*. The “Human Coders” columns reports the average of the IAA scores between the five human coders for each displayed label. The “Coders vs. Training” column displays the average IAA score of each human coder with the training dataset used for model training. The “LC” columns report the average of the Linguistic Classifier’s scores on the test folds during five-fold cross-validation. Scores are calculated loosely, meaning one code agrees with another code if it exactly matches or overlaps that other code, and both codes have the same label.

Label	Coders' IAA			Coders vs. Training			LC		
	Precision	Recall	F ₁	Precision	Recall	F ₁	Precision	Recall	F ₁
Gendered Pronoun	0.961	0.954	0.957	0.986	0.956	0.971	0.811	0.992	0.893
Gendered Role	0.784	0.804	0.780	0.803	0.833	0.811	0.740	0.808	0.773
Generalization	0.381	0.295	0.277	0.981	0.138	0.236	0.646	0.232	0.341

Table 5: Comparing human coders and text classification models’ performance for *Feminine*, *Masculine*, *Non-binary*, *Unknown*, and *Occupation*. The “Coders’ IAA” columns reports the average of the IAA scores between the five human coders for each displayed label. The “Coders vs. Training” column displays the average IAA score of each human coder with the training dataset used for model training. The “PNOC” columns report the average of the Person Name and Occupation Classifier’s scores on the test folds during five-fold cross-validation. Scores are calculated loosely, meaning one code agrees with another code if it exactly matches or overlaps that other code, and both codes have the same label. There were no *Non-binary*-coded text spans in the training dataset so no scores are reported for that code

Label	Coders' IAA			Coders vs. Training			Baseline PNOC			Cascade 1 PNOC		
	Precision	Recall	F ₁	Precision	Recall	F ₁	Precision	Recall	F ₁	Precision	Recall	F ₁
Feminine	0.604	0.601	0.597	0.992	0.682	0.807	0.636	0.646	0.641	0.487	0.675	0.566
Masculine	0.686	0.667	0.665	0.997	0.696	0.817	0.506	0.344	0.409	0.415	0.398	0.406
Non-binary	-	-	-	-	-	-	-	-	-	-	-	-
Unknown	0.677	0.725	0.678	0.996	0.800	0.879	0.589	0.636	0.611	0.591	0.464	0.520
Occupation	0.660	0.749	0.698	0.934	0.809	0.866	0.657	0.599	0.627	0.652	0.569	0.608