NORTHWESTERN UNIVERSITY

# Biases as Values: Evaluating Algorithms in Context

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Technology & Social Behavior

By

Mark Díaz

EVANSTON, ILLINOIS

June 2021

# Abstract

Biases as Values: Evaluating Algorithms in Context

Mark Díaz

This dissertation asks how researchers can create more equitable algorithmic systems. Ultimately, this thesis explores methods and implications of representing subjects of analysis in the design and evaluation of algorithmic systems. I also unpack how algorithmic tools measure and quantify human behavior, giving heed to the potential impacts of these systems on underrepresented communities. Building off of current work in HCI and algorithmic fairness, my research raises questions about how we can evaluate algorithms to understand the contexts and communities they serve best. The data sets and survey data used in this research are available at https://dataverse.harvard.edu/dataverse/algorithm-age-bias/. Because algorithmic tools are often created using data processed in similar ways (e.g., using one of a few common, publicly available data sets, or generating training annotations exclusively through crowd work platforms), these tools can fail to capture data and patterns that reflect underrepresented groups. The result can be unintended algorithmic bias. I use mixed methods– both quantitative and qualitative– to explore the broader social contexts in which algorithmic tools are applied and test methods of mitigating unintended algorithmic bias—one of which directly involves subjects of analysis in the creation and evaluation of an algorithmic tool.

In my first study I systematically analyze the outputs of popular sentiment models for age bias. I find a tendency for text to be rated more negatively when it references older age, then I develop an approach to removing bias rooted in training data. Taking older adults as a stakeholder group to prioritize in the face of age bias, I next solicit data annotations from older adults to evaluate model performance against their expertise on age and aging. A new model built from this data replicates age bias of similar magnitude to bias in my initial analysis and raises questions about the influence of annotator social identity and beliefs on model performance. In the final study I complement quantitative approaches to model assessment and turn to qualitative methods to evaluate model objective functions using direct

input from algorithm stakeholders. Ultimately, I argue that the development of ethical algorithmic tools must involve input from the very individuals who will be analyzed and impacted by system deployment.

# Acknowledgments

Many thanks to my advisor Darren Gergle for giving me the freedom to choose my research directions, learn how to answer questions that matter most to me, and become a scholar without compromising the convictions I hold as a queer, Black, and Latinx man. Additional thanks to Anne Marie Piper and Nick Diakopoulos for their guidance and always leaving me with thought-provoking questions. Thank you to Sheena Erete and Jessa Dickinson for introducing me to the exciting work of community-based research and keeping me motivated in my own projects when my motivation for research was waning. Much gratitude to friends, colleagues, and loved ones (Kima, Hillary, Halle, Anthony, Remington, Jeremy, Marie, Chris, Scott, and CollabLab) who have made me a more critical person and continue to provide indispensable distraction from work.

And finally, I'm indebted to my parents and family who fled authoritarian governments, came to the U.S., worked it out, and made it possible for me and my brothers to work it out.

# Table of Contents

Abstract 2

Acknowledgments 4

List of Tables 7

List of Figures 10

Chapter 1. Introduction 12

1.1. Algorithmic Social Bias & Methodological Approaches 14

1.2. Research Studies 16

Chapter 2. Background and Related Work 19

2.1. Social Implications of Bias 19

2.2. Technical Investigations of Bias 21

2.3. Emphasizing Marginalized Experience 23

2.4. Applying Computational Tools in Research 24

Chapter 3. Focusing on Data 26

3.1. Bias and Human Behavior 26

3.2. Study 1: Addressing Age-Related Bias in Sentiment Analysis 29

3.3. Phase 1: Explicit Encoding of Age 32

3.4. Phase 2: Implicit Encoding of Age 36

3.5. Phase 3: Addressing Age Bias via Training Data 41

3.6. Strategies for Addressing Bias 44

Chapter 4. Bringing Stakeholders into Data 47

4.1. Study 2A: Older Adult Input in Evaluation 49

4.2. Study 2B: Older Adult Input in Training 57

4.3. Annotator Analyses 65

Chapter 5. Responding to Social Bias 74

5.1. Implications of Age Bias 74

5.2. Removing or Preserving Bias 77

5.3. Testing 84

Chapter 6. Identifying and Evaluating Unknown Social Biases 87

6.1. Study 3: The Walk Score 89

6.2. Method 94

6.3. Findings 97

6.4. Walkability in Context 105

6.5. Walk Equity 111

Chapter 7. Data & Measurement 114

7.1. Mediated through Metrics 115

7.2. Expanding Stakeholder Inclusion 120

7.3. Improving Data Collection, Generation, and Analysis 124

Chapter 8. Conclusion 133

8.1. Contributions 135

Bibliography 138

Appendix 158

8.2. Demographic Survey 158

8.3. Aging Experience Survey 160

8.4. Aging Anxiety Survey 163

8.5. Age-Related Test Set 164

# List of Tables

3.1 The 15 sentiment analysis methods examined, and their corresponding type and validation data used when building the model. Validation data that is not social-media-based is predominantly based on movie or product reviews or news corpora. 32

3.2 Regression results for explicit age analysis. The results indicate the likelihood of a "positive" or "negative" outcome (rather than "neutral") given the variable input at the top of each column. The models include data from all sentiment analysis tools and are multinomial log-linear regressions, resulting in a model for positive sentiment and a model for negative sentiment. The reference categories are: neutral sentiment, "old" adjectives (i.e., "old" or "older"), lexicon-based approaches, and non-social-media validation data. Exponentiated coefficients (i.e., e^coef) provide relative risk (e.g., the sentiment analysis models were 1.66 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "older" adjective to a "younger" adjective"). Note: *$p$<0.05; **$p$<0.01 35

3.3 Individual regression results for explicit age analysis. The results from each sentiment analysis method were fit to a multinomial log-linear regression model, resulting in a model for positive sentiment and a model for negative sentiment for each sentiment analysis method. The reference categories for each model are: neutral sentiment and "old" adjectives. Coefficients that are not significant at $p$<0.05 are greyed out. Exponentiated coefficients (i.e. ecoef) provide effect sizes for relative risk (e.g. the EmoLex model was 3.18 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from "old" (or "older" or "oldest") to "young" (or "younger" or "youngest") holding all else constant. Note: *$p$<0.05; **$p$<0.01 The Sentiwordnet model (corpus-based) is not included because it did not classify any sentences "neutral." Instead, I used "negative" as the reference category for regression. The model was 4.121 times more likely to indicate positive for a "young" sentence compared to an "old" sentence ($p$<0.01, 95%CI: [2.390, 7.106]). 37

3.4 Details on the 10 GloVe models. The first part of the name references the source, the second part of the name gives the number of tokens contained in the source (e.g., 6B = 6 billion), and the third part of the name gives

the number of dimensions of the word vectors (e.g., 300D = 300-dimensional vectors for each word in the vocab). Further details at https://nlp.stanford.edu/projects/glove/ 38

3.5 Individual regression results for the implicit age analysis. The results from each sentiment analysis method were fit to a multinomial log- linear regression. The reference categories for each model are: neutral sentiment, and "control" adjectives. Exponentiated coefficients (i.e. e^coef) provide effect sizes for relative risk (e.g. the top right coefficient -0. the EmoLex model was 1.134 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "control" adjective to an "older" adjective as determined by the word embeddings. Note: *$p$<0.05; **$p$<0.01 40

3.6 The increase in likelihood that a "young" sentence will be classified as "positive" compared to its "old" counterpart. Training the model on the full, original dataset, a "young" sentence was 13.26% more likely to be "positive" compared to its "old" counterpart. There were 169 "old" and "young" sentence pairs. 43

3.7 Paired t-test results for the custom-trained classifiers. A likelihood above .50 produces a classification of "positive." 43

4.1 Test data set descriptions. 51

4.2 The break down of annotator age. 52

4.3 The break down of annotator gender. 52

4.4 The break down of annotator region of residence. 52

4.5 The break down of annotator race and ethnicity. In total there were 1,483 annotators. 52

4.6 The sentiment models and training data sets. 53

4.7 Each model's performance on the *Age-Related* test set. The McNemar test indicated a marginally significant difference between each model's performance on the *Age-Related* test set, $\chi^2 = 2.78$, $p$<0.10. The test set contained 296 sentences sourced from blogs authored by older adults. The *Original* model classified 199 inputs correctly and the *Age-Removed* model classified 190 inputs correctly. The models disagreed on 23 inputs. 55

4.8 Each model's performance on the *Non Age-Related* test set. The McNemar test indicated no significant difference between each model's performance on the test set, ($\chi^2 = 0.0$, $p$=1.0). The *Non Age-Related* test set contained 358 sentences sourced from Sentiment140. 56

4.9 The sentiment models and training data sets. 60

4.10 Each model's performance on the *Age-Related* test set. The *Age-Removed* model is shown for comparison. The test set contained 296 sentences sourced from blogs authored by older adults. The Older Adult model classified 195 examples correctly and erred significantly differently from both the *Original* model ($\chi^2 = 23.31$, $p$<0.0001) and the *Age-Removed* model ($\chi^2 = 9.63$, $p$<0.002). 64

4.11 Each model's performance on the *Non Age-Related* test set. The McNemar test indicated no significant difference between the *Original* model's performance and the *Older Adult* model's performance ($\chi^2 = 0.0$, $p$<1.0). There was also no significant difference in error patterns between the *Older Adult* model and the *Age-Removed* model ($\chi^2 = 0.500$, $p$=0.4795). The test set contained 358 sentences sourced from Sentiment140. The models disagreed on just one input. The *Age-Removed* model is shown for comparison. 65

4.12 The CrowdTruth agreement metrics. 66

4.13 The t-test results comparing anxiety group performance on the test sentences referencing *older* age indicate no significant difference between each group's tendency to annotate more positively or negatively ($p$<0.469, $M_{High}$-$M_{Low}$=-0.11, 95%CI= [-0.41, 0.19] SE=0.153). 68

4.14 The t-test results comparing anxiety group performance on the test sentences referencing *younger* age indicate no significant difference between each group's tendency to annotate more positively or negatively ($p$<0.554, $M_{High}$-$M_{Low}$=0.08, 95%CI= [-0.19, 0.36] SE=0.140). 69

4.15 The demographics of the High Age Anxiety Group that annotated the Age-Related test set sentences referencing older age. 70

4.16 The demographics of the Low Age Anxiety Group that annotated the Age-Related test set sentences referencing older age. 70

4.17 The annotation categories and their quality scores. 71

6.1 Neighborhood characterization of population demographics, income, and education according to data from the 2017 American Community Survey Estimates. 94

# List of Figures

3.1 A simplified flow chart showing steps in the model creation process. 26

3.2 Study 1 focuses on real-world applications of computational tools– specifically, real data inputs for analysis and their corresponding outputs. 29

4.1 Study 2 focuses on real-world applications of computational tools– specifically, real data inputs for analysis and their corresponding outputs. 50

4.2 Shown are hypothetical outputs of Model A and Model B, where red indicates incorrect classifications. The models rated 4 sentences incorrectly each, producing 60% accuracy. However there is no overlap in which examples they classified incorrectly. A significant McNemar test indicates the existence of many unique, non-overlapping errors in the outputs of each model. 54

4.3 The distribution of sentiment annotations provided by older adults. In total there were 1,483 annotators. 61

4.4 The distribution of "positive," "neutral," and "negative" annotations on the *Older Adult* training data subset featuring "old" and "young." 66

4.5 The distribution of "positive" and "negative" annotations on the *Original* training data subset featuring "old" and "young." 66

4.6 The distribution of annotator age anxiety. 67

4.7 The bar graph shows the average annotation for each anxiety group (High= -0.16, Low =-0.06). The results of the t-test indicate that there is no significant difference between each anxiety group's tendency to rate sentences more positively or negatively than the other. Thus, age anxiety does not predict annotation behavior. The difference between each group was just .10, which is small relative to the Likert scale unit step of 1. 69

4.8 The distribution of annotator agreement across all annotated examples. 71

6.1 Study 3 focuses on the evaluation of the variables defined and considered in a walkability algorithm. This process ultimately raises questions about a number of other stages in the model creation process. 88

6.2 Participant race and ethnicity. 95

6.3 Participant gender and age. 95

CHAPTER 1

# Introduction

Across the United States, the Transportation Security Administration (TSA) uses body scanning technology in airport security checkpoints as a way of detecting concealed weapons and foreign objects. As part of their operation, body scanners require operators to make a determination of a traveler's sex to manually select a corresponding option on the control dashboard. This action prompts the software to ignore breasts and penises when detecting concealed objects for women and men, respectively. While the scanners were introduced to bolster efficient and rigorous security processes for all travelers, one unintended consequence was a more negative and complicated screening process for many trans* and gender nonconforming (GNC) people 1. Individuals whose bodies challenge normative notions of gender, such as trans* and GNC individuals, are routinely flagged as anomalies, causing experiences of humiliation, discomfort, and fear (Cascio, 2018). In addition, the inability of body scanners and software to properly detect concealed objects on the bodies of many trans* and GNC individuals forcibly subjects them to lengthier, unnecessary security screening procedures, such as pat downs, that may cause further discomfort. With the population of trans* and GNC individuals in the United States estimated to be less than 1% (Flores, A.R., Herman, J.L., Gates, G.J., & Brown, T.N.T, 2016) it is unlikely that the design and evaluation of body scanners and concealed object detection software was informed much, if at all, by the experiences of trans* and GNC people. At that, the political consequences of formally recognizing trans and GNC inequality likely informs TSA's inattention to the issue. A more inclusive design process may have allowed designers to better evaluate pitfalls in how body scanning systems interpret bodies and how the systems might be designed to more fully serve their contexts of use.

Beyond airport body scanners, algorithms play an increasingly significant role in the design and deployment of sociotechnical systems that touch upon nearly every aspect of daily life. As researchers and engineers seek to refine algorithmic performance, these systems have been and continue to be employed for high-stakes analyses ranging from criminal justice issues (Berk et al., 2018) to day-to-day municipal processes in cities across the world (Holdren & Lander, 2016). Experts have started to point out that the promises of efficiency and productivity that data-driven systems bring

1Trans* individuals are those whose gender identity differs from the gender they were assigned at birth, while GNC individuals are those whose gender expression does not conform to societal expectations

hinge on processes that can also further marginalize and disenfranchise underserved communities (Eubanks, 2018; O'Neil, 2017). Researchers in Human-Computer Interaction (HCI), have highlighted that algorithmic bias can emerge from discrepancies between contexts from which data is procured, the contexts algorithmic tools are intended for, and the contexts in which algorithmic tools are ultimately employed (Nissenbaum, 2001). In response to ethical concerns, researchers are coming up with new ways to audit these systems (Diakopoulos, 2015), placing focus on various issues including optimization, algorithmic bias (or fairness), interpretability, and user adoption. In Data Feminism, D'Ignazio and Klein (2020) highlight a need to address specific issues in the ways individuals and communities are too often "uncounted, undercounted, and silenced" in current data science practices (D'Ignazio & Klein, 2020). The work I present in this dissertation speaks directly to their call to action for researchers and data scientists to "bring the bodies back" to data science practice. They advocate for including subjects of analysis, particularly those who are underrepresented, in the pipeline of building and evaluating algorithmic technologies. I respond to their call by tracing the roots of age bias in algorithms and experiment with new approaches to soliciting input from subjects of analysis in various stages of algorithmic design. I lay groundwork for building upon this work through qualitative field methods as an effort to improve how researchers can uncover biases, identify algorithm limitations, as well as improve equity in data representation.

Ultimately, this thesis explores methods and implications of representing subjects of analysis in the design and evaluation of algorithmic systems. I argue that designing algorithmic tools for appropriate analyses of underrepresented communities can be improved by soliciting input from the very individuals who will be analyzed and impacted by system use. Building off of current work on algorithmic transparency and assessing ethical uses of algorithmic tools (Gebru et al., 2018; Hanna et al., 2020; Mitchell et al., 2019), my research raises questions about how we can evaluate and validate algorithms to understand the contexts and communities they serve best.

Toward exploring productive ways of incorporating the voices of underrepresented groups in data work, I unpack how algorithmic tools measure and quantify human behavior, giving heed to the potential impacts of these systems on underrepresented communities. Decisions around the deployment and use of algorithms and algorithmic systems need to be critically considered in their specific contexts of use. By "contexts of use" I refer to the types of analyses and use cases for which algorithmic systems are employed as well as connected stakeholders. Stakeholders include not only direct end users, who might be analysts, consumers, or government employees, but also individuals and communities impacted by decisions or processes shaped by system outputs. I also pay particular attention to a given model's fitness

for a context of use. I define a model as fit for a context of use if it produces outputs that are validated against stakeholder values and needs, particularly stakeholders who are underrepresented or marginalized. A fit algorithmic system may include technological components that produce socially biased results but decision-making processes involving subject matter experts who can provide supplemental knowledge or interpretations that consider the broader sociotechnical system. For example, U.S. census data does not reliably report homeless and housing unstable individuals which complicates procurement and disbursement of needed resources in municipalities (Wright & Devine, 1992, 1995). While a predictive system trained primarily using census data is at risk of propagating biases associated with this error, knowledge from experts working with homeless and housing unstable communities may be able to balance these biases or inform policy decisions that account for predictive limitations.

Incorporating underrepresented voices in data work is a departure from typical practices in research and data science, which often process data in similar ways (e.g., using one of a few common, publicly available data sets, or generating training data exclusively through crowd work platforms). As a result, algorithms can fail to capture data and patterns that reflect underrepresented groups. Algorithmic tools intended to be used broadly or universally, in particular, can fall victim to patterns represented in the limited data they are built upon, such as in the case of algorithmic walkability metrics failing to account for accessibility needs (Prescott, 2014). The result can be unintended bias in the outputs algorithmic tools and systems produce. In this dissertation I investigate and develop methods that consult underrepresented voices as one avenue to address unintended bias.

### 1.1. Algorithmic Social Bias & Methodological Approaches

In this dissertation, I focus on *algorithmic social bias*—that is, algorithmic bias that reflects negative mainstream or stereotypical attitudes about groups of people. This kind of algorithmic bias may not be inherently bad for a system to exhibit. Opinion mining algorithms, for example, are meant to reflect the opinions of a specific population, even if those opinions are controversial. Regardless, it is crucial that designers and users of algorithmic tools be made privy to the sources of biases that a system exhibits as well as the implications of those biases in intended contexts of use so that quantitative outputs can be interpreted and used responsibly. As sociologists Espeland and Stevens point out, "quantification is fundamentally social – an artifact of human action, imagination, ambition" (Espeland & Stevens, 2008). The core of my work is driven by this assertion and culminates in an examination of ways in which members of an underrepresented group—namely older adults—can be incorporated into the creation and evaluation

of an algorithmic tool. This dissertation contributes a method and process for understanding bias in algorithmic tools and their suitability for analyzing specific populations.

From both quantitative and qualitative perspectives, there are challenges regarding how to design, implement, and evaluate algorithmic systems with respect to underrepresented communities, many of which have unique needs and are disproportionately vulnerable to adverse effects (Eubanks, 2018). Critical Algorithm Studies (CAS) has painted a vivid picture of the ways in which algorithms interact with society, particularly underrepresented groups, and I draw on CAS to frame the broader impacts of algorithmic technologies. Many scholars in CAS have examined algorithmic bias, namely the ways in which automated systems propagate social biases against certain groups and individuals (Boyd et al., 2014; Noble, 2018). Prior work analyzes algorithmic bias in search engines, surveillance systems, and social media (L. D. Introna & Nissenbaum, 2000; L. Introna & Wood, 2004). For example, Taina Bucher has analyzed the ways in which social media algorithms control the visibility of different content and, therefore, users' ability to "see and be seen" (Bucher, 2012). Other research in this area has framed algorithmic bias as an instance of technology embodying social, ethical, and political values (Nissenbaum, 2001). I also draw on CAS to contextualize quantitative investigations of algorithmic bias. I inform my quantitative approach with CAS's emphasis on centering members from underrepresented communities to consider a specific case study of how underrepresented voices can be used to shape and evaluate algorithm design.

To translate insights from CAS into the "how" of designing and testing algorithmic technologies, I turn to HCI and Fairness, Accountability, and Transparency (FAccT) literature. While scholars in CAS shed light on the intersections between algorithms and culture and power, researchers in HCI and the growing FAccT communities have emphasized empirical assessments and methods to address algorithmic bias and algorithmic fairness in machine learning. Work in HCI and FAccT has entailed developing frameworks and technical approaches for auditing algorithmic tools as well as probing end users' perceptions of algorithmic fairness and justice (Binns et al., 2018; Woodruff et al., 2018). This line of research has contributed to important conversations about how to identify the extent to which different kinds of biases are problematic for specific application domains. Challenges surrounding algorithmic fairness emerge in processes that involve human touch-points with algorithm creation and evaluation. These processes include the use of crowd workers to generate training and testing data (Sen et al., 2015) and algorithmic problem formulation (Martin Jr et al., 2020). HCI literature highlights ongoing work tracing the emergence of bias (Friedman & Nissenbaum, 1996) as the ways in which algorithmic outputs are interpreted by end users (Eslami et al., 2015).

To conversations in FAccT on empirical measurements of bias, I contribute evaluations of in- and out-of scope uses of algorithmic tools using both qualitative and quantitative approaches, as well as methods of modifying existing model architectures with stakeholder input to make them more appropriate for novel contexts. I aim to create dialog that interrogates when and why different methods might support or inhibit research goals. A challenge in the design of these tools is how to align the definitions of problems and concepts that an algorithm is intended to learn with those of different stakeholders. If an algorithmic metric intended to measure walkability, for example, is built on definitions of walking behavior that differ from the behavior of a particular population, algorithmic outputs may be limited or unhelpful. In this work I unpack these alignments and misalignments taking inspiration from values-based approaches. Values-based approaches have proven helpful for evaluating stakeholder priorities and conceptual definitions to help ensure inclusive design practices. Value Sensitive Design was originally outlined by Friedman et al. (2002) as a theoretically-grounded approach that accounts for human values throughout the design of technological and information systems. For underserved and historically marginalized populations, in particular, a focus on values helps elicit stakeholder values as well as the differing logics with which various stakeholders understand the same values and concepts (Voida et al., 2014).

### 1.2. Research Studies

Through several studies I pursue the question of how researchers can develop more equitable algorithmic systems. I argue for the importance of consulting with and expanding the involvement of underrepresented and marginalized communities in the design and evaluation of algorithmic technologies– both as a matter of improving how we deal with social bias and as a matter of shifting societal power imbalances. Up to this point, scholars in algorithm design have dedicated scholarship to understanding marginalized populations (Hanna et al., 2020; Woodruff et al., 2018) and have applied participatory approaches to the algorithm design pipeline Martin Jr et al. (2020), Zhu et al. (2018). I build on these investigations to modify existing machine learning approaches and then test them specifically against underrepresented viewpoints, with the goal of amplifying marginalized voices. These studies both evaluate existing algorithmic tools to understand how they characterize subjects of analysis as well as explore methods of addressing unintended social bias. The first two research studies in this dissertation consider the capabilities and limits of algorithms in measuring human behavior, particularly in analyses of underrepresented populations– the second of which involves members of an underrepresented population in the creation and evaluation of the model. The third

research study pivots away from quantitative evaluation of bias and instead proposes qualitative methodology as a valuable tool to validate algorithmic design choices.

A first step toward building an equitable system is to identify aspects or features that are *inequitable*. In the context of algorithmic systems, this entails both identifying algorithmic social bias, as well as bolstering the techniques available to identify new and evolving forms of social bias. Because computational systems feature a wide range of biases, social and otherwise, researchers must determine whether the biases exhibited by a chosen algorithmic tool are significantly detrimental to the analyses at hand. At that, researchers must be consider differential negative impacts of biases on stakeholders of decisions that hinge on analysis results. For example, a language model that systematically rates references to a protected group as more negative than others is potentially more harmful than a language model that artificially biases *all* outputs to be more negative. At the highest level, this dissertation asks how researchers can create more equitable algorithmic systems by leveraging the knowledge and experience of stakeholders who stand to be most impacted by system deployment. The primary research questions addressed by the following studies are: 1. *How can researchers evaluate algorithmic social bias for their research contexts?* 2. *How do algorithmic tools characterize attitudes and experiences—particularly those of underrepresented populations— in model outputs?* 3. *How might unintended social bias be mitigated– ideally in ways that incorporate knowledge from underrepresented and marginalized populations?*

Study 1 takes a quantitative approach to understanding age bias with respect to sentiment analysis algorithms and old age. This study homes in on a specific case study of how algorithmic tools characterize older age and traces roots of age bias in model data. This study highlights potential issues with the application of algorithmic tools that are intended to broadly analyze human behavior, but that may not be sensitive to or evaluated with underrepresented groups in mind. In a systematic analysis of 15 sentiment analysis tools, which are intended to measure human sentiment and attitude in text, I, together with my colleagues (Dıaz et al., 2018) show bias in the treatment of age-related terms. The results demonstrate a divergence between how sentiment analysis tools characterize older adults and how older adults characterize themselves, according to existing literature on age and aging.

Although Study 1 is successful in demonstrating a method to mitigate age bias in our research context, I use Study 2 to show how researchers might better validate algorithms by using stakeholder input in model testing and training. Responding to Sen et al. (2015)'s call for algorithms to be evaluated against specific communities (Sen et al., 2015), I solicit input from older adults. In doing so, I involve older adults as a matter of data representation. By using input

from older adults in testing and training sentiment models, I seek to leverage their nuanced experience with age and capture their viewpoints and ground truth. In this study I assess sentiment model accuracy against older adult ground truth, and train and test a custom model using older adult data annotations.

The third and final study in this work qualitatively explores an algorithm in context by focusing on the Walk Score, a patented algorithm for quantifying the walkability of a chosen geographic area. This study draws on interview methods to explore alignments and misalignments between how the Walk Score algorithm defines walkability and neighborhood life and how neighborhood residents define their own experiences. The study takes a value-sensitive approach to outlining values designed into the Walk Score algorithm, contrasting them with those that emerge in participant interviews. The work highlights the limits of algorithmic metrics in capturing subjective experience. While the Walk Score is suitable in the context of consumer real estate searches, its inability to capture variation in subjective experiences of walking problematizes its suitability for use in population-level analyses, such as in measurements of community-wide physical activity (Jones, 2010). Study 3 helps raise questions about appropriate and inappropriate applications of not just algorithms, but also the limitations of quantification for measuring marginal experience.

CHAPTER 2

# Background and Related Work

I begin my investigations of algorithm design by drawing from critical social scholarship, computer science, and HCI. Each of these interrelated areas provides necessary perspectives that allow me to evaluate algorithmic technologies in terms of both their broader social implications as well as their technical design specifications. Whereas critical social research places emphasis on understanding the macro-level societal context of algorithmic technologies and technical investigations often place emphasis on technical solutions to discovering and mitigating unintended bias, HCI sheds light on how individuals use and understand algorithmic technologies, rounding out a sociotechnical lens. Although the present dissertation largely focuses on methods of parsing sources of unintended algorithmic bias in data, I contribute to ongoing discussions in HCI that interrogate how algorithmic systems are used in the real world. This work ranges from investigations of how individuals perceive the justice of algorithmic decisions (Binns et al., 2018) to considering the needs of practitioners who rely on algorithmic technologies for doing their work (Holstein et al., 2019). Understanding an algorithm's technical design, its connections to human stakeholders and, ultimately, real-world social impacts is necessary for comprehensively evaluating algorithmic performance. In this dissertation I consider typical metrics used to evaluate algorithms, such as accuracy and F1 score. However, I also consider how algorithm designers should solicit ground truth data for building and testing, particularly when ground truth might be established in a multitude of ways.

## 2.1. Social Implications of Bias

With the goal of understanding how algorithmic systems characterize human behavior, the limits thereof, and how underrepresented populations might be better incorporated into algorithm design, I begin with literature that frames the broader social impacts of algorithmic technologies. Research in critical algorithm studies (CAS), sociology, science and technology studies, communication, and legal studies characterize the ways in which algorithms operate and influence social life. Crucially, work on algorithms in these fields analyzes technology in relation to power hierarchies in society such as those of race, gender, and geography. It is through this work that algorithms' connections to social inequity become salient.

Scholars focused on the sociotechnical nature of algorithms have situated them as influential forces that shape and are shaped by social, political, and cultural life (Gillespie, 2014). Looking beyond the technical specifications of algorithmic systems, scholars have highlighted the ways in which humans interact with and are acted on by algorithms, defining algorithmic bias as, "systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman & Nissenbaum, 1996). Although the term "bias" has many different uses and does not always carry a negative connotation, I note that scholars pursuing sociotechnical work often use "bias" to refer to *social* biases and discrimination.

Algorithms have become information gatekeepers in social media systems (Bucher, 2012) and surveillance technologies, such as facial recognition (L. Introna & Wood, 2004), as well as arbiters in the provision of social services (Eubanks, 2018). Importantly, they are artifacts with embedded values and biases that can reinforce and expand systems of oppression (Noble, 2018). Many scholars in critical algorithm studies have examined algorithmic bias by looking to the outputs of automated systems and the ways they unfairly discriminate against certain groups and individuals in favor of others (Diakopoulos, 2015; Noble, 2018; Thebault-Spieker et al., 2017). For example, Taina Bucher has analyzed the ways in which social media algorithms control the visibility of different content and, therefore, users' ability to "see and be seen" (Bucher, 2012). Similarly, Introna and Nissenbaum describe the ways that biased search engines diminish access to information as well as individuals' abilities "to be seen and heard" (L. D. Introna & Nissenbaum, 2000). In their foundational contributions to Value Sensitive Design (VSD), Friedman and Nissenbaum have framed computer systems as expressions of social, ethical, and political values (Friedman, 1997; Nissenbaum, 2001), and further work has focused on understanding the sources of bias and identifying ways to diminish it (Caliskan et al., 2017; Dixon et al., 2018).

There is also growing interest in issues of social justice specifically in HCI, as evidenced by new frameworks and agendas (Bardzell, 2010; Dombrowski et al., 2016; Irani et al., 2010; Lazar et al., 2017) that attempt to shift power balances between researchers, society, and marginalized groups. These frameworks tackle diverse domains but converge on several points. Given that algorithmic technologies can play a role in both reinforcing oppressive power structures, it is necessary to interrogate the implications of algorithmic technologies for underrepresented and marginalized populations. A growing body of work draws from Value Sensitive Design (VSD) in HCI and calls attention to algorithmic bias as an instance of technology embodying social, ethical, and political values (Nissenbaum, 2001). The primary takeaway is that science, technology, and design are not neutral or valueless; rather, they perpetuate distinct

viewpoints or ways of thinking. This fundamental idea reframes bias in algorithm performance from a quantitative observation to a phenomenon that can be readily linked to the social contexts in which algorithms and their underlying data originate. In response, a number of scholars have made calls for efforts to regulate algorithmic systems and for designers to create more ethical systems to ensure the survival of underrepresented communities (Crawford & Schultz, 2014; Diakopoulos, 2014). Algorithm designers must contend with whose values are in alignment with particular algorithmic systems, and whose values are not.

## 2.2. Technical Investigations of Bias

To translate critical social insights into actionable algorithm design and evaluation practices, I turn to technical and modeling investigations largely stemming from HCI and Fairness, Accountability, and Transparency (FAccT) literatures. Whereas critical social research provides rich qualitative accounts and case studies of the impacts of algorithmic systems, technical investigations provide complementary empirical investigations that both measure bias and explore technical solutions. These works pick up on the socially-driven investigations of algorithmic bias through building, testing, and auditing algorithmic technologies (Celis et al., 2019; Friedler et al., 2019; Madaan et al., 2018). These analyses are expressly focused on the underlying mechanisms that drive bias in algorithms. With nuanced understandings of underlying mechanisms, researchers are able to experiment with data collection and data processing in service of meeting desired fairness criteria. Empirical investigations of the impacts of algorithmic bias provide important methods for evaluating algorithms that use assessments beyond accuracy and F1 score, which have traditionally been the primary metrics for determining algorithmic fitness.

Recent research has both characterized the presence of unintended bias in algorithmic systems as well as laid out approaches and frameworks to prevent it. Žliobaitė (2017) points out potential roots of social bias in model design, including data pre-processing, model post-processing, and model regularization (also referred to as in-processing by (Hajian & Domingo-Ferrer, 2012; Veale & Binns, 2017)). The studies of this dissertation work focus primarily on model design. As a component of model design, data pre-processing in algorithmic bias assessments involves modifying training data to remove or reduce differences across protected groups. Kamiran and Calders (2009) took this approach in mitigating bias in a data set used for training a credit approval system, and repeated a similar process using both synthetic and real-world data sets (Kamiran et al., 2013). Model post-processing involves creating a model and modifying its architecture to meet some non-discrimination criterion, such as by removing specific leaves of a decision tree. For example, Calders and Verwer (2010) experimented with this approach by modifying the probability

distribution of the attributes of a protected class in a Naive Bayes classifier used to predict individuals' income levels. Like model post-processing, model regularization constrains outputs to specific non-discrimination criteria. However in model regularization this is done the model training phase. Using decision trees as an example once again, this could involve an adjustment to splitting criteria while the model is being built (rather than removing leaves after the model has already been trained). Applying a regularizer to probabilistic models, Kamishima et al. (2012) developed a technique to force model predictions' independence from sensitive features.

Toolkits for algorithmic model design and reporting have also emerged as valuable means for expanding social bias evaluation. Building from Gebru et al. (2018)'s proposal for a more standardized way of documenting data sets, Mitchell et al. (2019) created informational templates in support of transparency and responsible use to help designers of algorithmic systems clarify intended application contexts. The open-source Aequitas toolkit, created by researchers at the University of Chicago, puts forth tools for researchers and designers to audit algorithms for various kinds of unintended bias (Saleiro et al., 2018). Work using these tools highlights the importance of determining in- and out-of-scope applications of algorithmic tools, such as automated facial recognition. While not always an explicit research goal, technical investigations emphasize the role of computer scientists in shaping the outcomes and applications of algorithmic technologies. I extend Green's (2018) assertion of computer scientists in criminal justice reform to all algorithm designers that we "must abandon naive notions of neutrality and recognize themselves [sic] as participating in normative and political constructions of society." My work contributes to these conversations by exploring methods of validating algorithmic technologies against specific contexts of use. Part of this exploration entails collecting direct stakeholder input to trace and shape social bias in algorithm outputs. I also discuss model performance in relation to computer scientists' role in problem formulation and data selection. Unlike prior empirical work on algorithmic social bias, the approach I take ultimately *treats algorithmic bias as an artifact to shape and evaluate for specific contexts* rather than an artifact to remove completely.

In order to be effective, algorithmic technologies must not only be fair, but also be usable, understandable, and easy to integrate into existing social and collaborative dynamics. In a lab study that involved making decisions about splitting roommate chores, M. K. Lee and Baykal (2017) found that participants determined an algorithmic tool to be less fair than decisions made through group discussion. In addition to equitable distribution of tasks, participants considered group member preferences and reasons for compromise in assessing the fairness of the ultimate decision. The researchers posit that algorithmic systems "should account for social and altruistic behaviors that may be difficult

to define in mathematical terms". Veale et al. (2018) specifically conducted an interview study with machine learning practitioners to understand their needs in implementing fairer algorithmic technologies. Among issues they highlighted is algorithmic authority, noting that, "prioritised lists or options, are understudied in relation to how these affect the mental models constructed by those using these systems day-to-day." These studies highlight the broader need for algorithmic technologies to both accurately present fair data as well as do so in a way that supports responsible interpretations and decision-making on the part of end users.

### 2.3. Emphasizing Marginalized Experience

As I've described, my primary focus is unintended social bias that emerges in computational technologies. Indeed researchers in data science and machine learning have dug into various methods of identifying and measuring how unwanted social bias in data propagates through algorithmic processes. An important component of serving underrepresented communities in technology design is understanding how algorithmic systems represent and are responsive to their lived experiences. Humans possess biases that can benefit some groups over others, and a sociotechnical framing allows us to acknowledge that technological systems are shaped by these biases. Tarleton Gillespie outlines dimensions of what he names *public relevance algorithms*, or algorithms that, "select what is most relevant from a corpus of data composed of traces of our activities, preferences, and expressions" (Gillespie, 2014). These activities, preferences, and expressions are intricately related to social identity, meaning that social behavior directly shapes the swaths of data from which algorithms interpret or learn. This is perhaps most apparent in algorithms that seek to infer gender and race from user data (Hamidi et al., 2018; Keyes, 2018). However, even when social identity is not an explicit feature of analysis, patterns intimately related to social identity can be learned by algorithmic systems, such as implicit associations with race, gender, or age (Caliskan et al., 2017). A central thread of this dissertation work is an interrogation of the limits of what algorithmic systems do and do not capture about human experience. In other words, whose behaviors and attitudes are taken up by algorithms and reproduced in system outputs. Algorithmic systems are increasingly managing products and services aimed at broad, diverse audiences (Schrock, 2018; Vlachokyriakos et al., 2016). Thus, it is important to ensure that algorithmic systems be designed in ways that are responsive to various stakeholders– whether they be direct end users or subjects of analysis.

In this work I explore challenges in using algorithmic tools to measure subjective experience and highlight that the process of operationalizing human experience, such as walkability, in an algorithm involves assumptions about humans' behaviors and preferences. In the case of the Walk Score, the assumed user is relatively able-bodied, values

the particular amenity categories included in the algorithm, and has a particular distance threshold for walking before opting to choose another mode of travel. Assumptions about end users and populations of study are inevitable; however, researchers using the Walk Score have little indication of the assumptions underlying algorithmic measurements and outputs. Assumptions about algorithms' users and contexts of use are an issue that Mitchell et al. (2019) begin to address with their framework for model reporting. Misalignments between assumptions made by algorithm designers and those made by end users of algorithms raise concerns about the validity of algorithmic outputs and potential bias that might stem from applying them universally across a wide range of contexts (Mittelstadt et al., 2016).

### 2.4. Applying Computational Tools in Research

My primary site of inquiry is the application of computational tools to study underrepresented populations, namely older adults. This focus stems from research I initially conducted with colleagues that involved scraping a data set of over 200,000 blog posts written by older adults. These posts were published over the course of 12 years and chronicled older adults' experiences with older age and aging. While the posts covered a wide array of topics, persistent themes included age discrimination and ageism.

Initial qualitative work focused on understanding the role of blogging platforms in facilitating discussion that challenged mainstream beliefs and negative stereotypes about aging. I turned to computational techniques to leverage the wealth of data and conducted exploratory analyses involving text analytics. One such focus was the sentiment of blog posts, particularly in relation to how older adults described aging experiences. Puzzling outputs in my exploratory probe led me to consider the extent to which the positive and negative associations with aging that older adults espoused might be inconsistent with positive and negative associations with aging encoded into the computational tools I was using. As an example, one passage that was rated very negatively by a language model I used was the following:

> "*When we get old, we are more likely than in younger years to be diagnosed with terrible diseases sometimes treatable, other times not so much. I like knowing that if I become terminally ill and if my life, in that circumstance, becomes too painful or otherwise unbearable, I will not be required by law to su er.*"

The author reflects on older age in a way that, while serious in tone, ends on a note of ease or perhaps peace of mind. In addition, the highly negative assessment by the language model wholly misses the rich complexity I interpreted through my own reading. I did not expect a quantitative measure of emotion for such a passage to represent all of its nuanced facets; however, the assessment raised questions about how content describing older age might be evaluated

by a slew of computational tools used daily by researchers and analysts. Unsure whether the outputs I saw stood as anomalies, I turned to the computational tools themselves in order to investigate the possibility of consistent age bias. The first study of this dissertation outlines the application of computational tools to study older adults as a relevant case to understand the implications of age bias in data. The work considers age bias across a range of natural language processing tools and specifically considers the presence of human social biases in underlying data.

CHAPTER 3

# Focusing on Data

To begin my exploration of social bias in algorithms, I focus on data– the testing data used to validate tools, the training data used to build algorithmic tools, and the sources from which this data comes. Data plays a necessary role in various steps of the model creation pipeline (3.1 shows a simplified process). Importantly, every step of the model development and implementation pipeline is touched by humans and therefore, an entry point for potential social bias. In addition to biases embedded in data and its sources are biases expressed by engineers and researchers who make decisions about which data to use and which data to ignore. After establishing the importance of underlying model data and its relationship to bias, I take a deep dive into identifying and measuring age bias in sentiment models by carefully constructing testing and training data sets.

## 3.1. Bias and Human Behavior

Human social biases shape both what becomes encoded in data sets used to create algorithms as well as how end-users make sense of algorithmic systems (DeVito et al., 2017; Eslami et al., 2016; Eslami et al., 2015). User inputs, for example, can affect bias in system operation within a sociotechnical context. Liao et al. (2016) examined the ways that Twitter users with minority opinions used design features such as the "retweet" button to amplify their views and introduce bias. Other researchers have attempted to systematically sort out user bias (e.g., the keywords users choose in a web search) from bias introduced by a system (Kulshrestha et al., 2017). Attempts to reduce bias, however, are complicated by the interplay between users and systems. Green and Chen (2019), for example, demonstrated how individuals can reintroduce social biases that have been purposefully removed from quantitative risk assessments, complicating attempts to produce fair outcomes. A critical component that influences algorithm behavior and the


Figure 3.1. A simplified flow chart showing steps in the model creation process.

outputs that users engage with is the underlying data that algorithms learn from in order to predict and perform. Underlying data and corresponding outputs are the initial sites of my inquiry.

As part of the larger discussion on algorithmic bias, recent work has begun to analyze the design and underlying mechanisms of algorithms that contribute to and exacerbate bias, with a call for more empirical studies (Kitchin, 2017). Nissenbaum (2001) stated that, "Fastidious attention to the before-and-after picture, however richly painted, is not enough". Instead, she states that what engineers and computer scientists can contribute to the field is "a fine-grained understanding of systems– even down to the gritty details of architecture, algorithm, [and] code," as these are essential to "explaining the social, ethical, and political dimensions of information technologies". Some researchers have directly manipulated open-source algorithms to reveal the extent of structural biases (Johnson et al., 2017). However, given that many algorithms are proprietary, researchers have also attempted to decipher algorithms by interpreting output while varying inputs (Chen et al., 2015; Diakopoulos, 2014)– a technique I also employ. My work takes a deep dive into the data used to train and validate algorithmic technologies, interrogating where these data originate and how data provenance can impact resulting system performance.

Intuitively, the relevance and number of data points in a data set are important considerations for their use in the design of a given algorithm. For example, a data set with too few data points, may not provide enough examples for an algorithm to identify important or accurate patterns. Additionally, in supervised learning tasks data *annotations* are also critical. Annotations can be collected or generated in a number of ways, but are typically provided by researchers or crowd workers. Supervised learning processes require the assembly of a training data set as well as annotations for each data point in the data set. A learning algorithm is able to detect patterns across annotations in relation to the training data such that, when presented with novel data to analyze, it can produce a likely prediction. In the case of sentiment analysis a sentiment model might be built from a data set containing thousands or millions of sentences, each with an annotation of "positive", "neutral", or "negative". Based on similarities among sentences in each annotation group, the sentiment model will predict the group that a novel sentence most resembles (i.e., "positive", "neutral", or "negative").

Ultimately, where data comes from and how it is annotated has significant influence over the resulting algorithm's accuracy, performance, and adaptability to different application contexts. In the first two studies of this work, I frame annotations as individual interpretations of annotators that are not objective or universally applicable. That is, the way individuals annotate data will be informed by their values and perspectives. Not only must researchers and engineers

carefully consider which data are collected for training processes, but they must also carefully consider whether the annotations align with algorithmic design goals. For example, in a study by Patton et al. (2019) analyzing differences in annotation behavior between social work graduate students and community members trained in data annotation, community members more often annotated social media interactions between gang members as aggressive. They posit that community members looked to different features of the social media interactions to indicate aggression compared with the social work graduate students, who were less familiar with the social context in which the social media interactions were taking place. This is not to say that social work graduate students necessarily produce poorer quality annotations; however, the success of using aggression detection to inform social interventions may depend significantly on whose interpretive lens is privileged in the annotation process. Understanding differences in annotation behavior and their influence on subsequent algorithmic performance, then, emerges as a necessary site of analysis.

Because of the convenience it affords in obtaining sufficiently large, annotated data sets, data used to train machine learning classifiers is often crowd-sourced. However, Patton's research complements prior work showing that different gold standard data sets produce different algorithmic outputs (Sen et al., 2015). Indeed a range of scholarly work demonstrates that the values and social biases that algorithms capture are related to the data upon which they are built, including the individuals that produce or annotate them (Binns et al., 2017; Caliskan et al., 2017; Frey et al., 2018). Sen et al. (2015) call for algorithms to be evaluated based on how well they work for specified communities– a stance I take in this work. Differences in data set characteristics raise the question of how data sets and annotations from different sources influence algorithmic performance. I frame community members' annotations as artifacts of their interpretive lens, which is informed by lived experience and familiarity with the contexts from which training data are extracted, such as in the case of community member annotators in Patton et al. (2019)'s work. It is precisely here that I begin my work to parse how biases become encoded in data sets and propagate through algorithm performance.

Study 1 grapples with measuring and mitigating age bias encoded in language models through a systematic analysis of popular models used to analyze sentiment in text[1]. In order to compare how data collection and annotation practices lead to different encoded biases, a baseline must be established for comparison. In addition to studying model architecture in relation to social bias, specifically age discrimination, the analyses culminate in an exercise of isolating a training data subset that significantly produces age-related bias. Ultimately, the work 1) illustrates how social biases propagate from train data to model performance, 2) outlines a simple method to detect and remove bias, and

[1]This work was originally published by Dıaz et al. (2018) at the ACM Conference on Human Factors in Computing

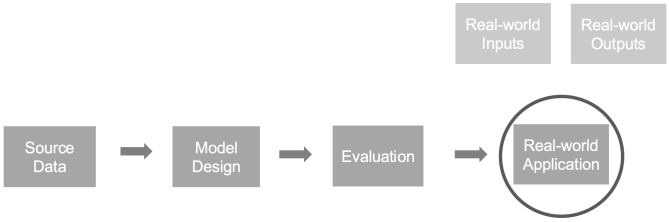


Figure 3.2. Study 1 focuses on real-world applications of computational tools– specifically, real data inputs for analysis and their corresponding outputs.

3) highlights complications with respect to collecting and manipulating training data with respect to underrepresented and marginalized populations.

### 3.2. Study 1: Addressing Age-Related Bias in Sentiment Analysis

I focus on age-related social bias in sentiment analysis as a case of using computational, algorithmic tools to study underrepresented attitudes and opinions. Specifically, I consider how age bias manifests in language models and its relationship to ageism in society. The analyses in this study focus on real-world applications of computational tools– specifically, real data inputs for analysis and their corresponding outputs (see 3.2).

Although ageism was identified several decades ago (Butler, 1969), negative attitudes and stereotypes about growing older are only recently receiving worldwide attention. The World Health Organization has recently called for a "global campaign to combat ageism," given the association between negative views about aging and decreased health and longevity (Officer et al., 2016). Age discrimination and age bias are topics that have received attention within HCI, in particular, where research has highlighted the ways in which researchers and designers tend to treat aging as a "problem" with technology as the solution, rather than viewing aging as a complex and natural part of the lifespan (Vines et al., 2015). To help counter age-related stereotypes around technology use, some researcher in HCI have reframed older age by emphasizing cases of older adults going online to actively create and share content (Brewer et al., 2016; Brewer & Piper, 2016; Harley & Fitzpatrick, 2009), learn to program (Guo, 2017), and even form social movements around ageism (Lazar et al., 2017).

There is a growing awareness of age discrimination worldwide (e.g., (Lahey, 2010; Officer et al., 2016)), and age-related bias in particular has not been studied with regard to popular sentiment analysis tools that are used to make strategic decisions about products, politics, finances, social services and employment (Diakopoulos, 2014; Kuncel et al., 2014; Miller & Record, 2017; Pasquale, 2015). Nor has there been much work specifically aimed at addressing

or reducing age bias in algorithms. Potential underlying bias around age has implications for the appropriateness of these tools in contexts where attitudes towards age matter, as well as the ways that subtle forms of age discrimination manifest in technologies that pervade everyday life.

It is important to understand the impact of these techniques and potential bias within a particular topic of study (Sen et al., 2015). In the case of age-related bias, automated methods of opinion polling surrounding issues related to older age may falsely report more negative attitudes toward political issues or financial investments regarding age-related concerns, such as Medicare and Social Security. Though bias may stem from many sources, I bring particular attention to addressing social bias rooted in training data. As previously discussed, the output of algorithms or machine learning models is largely dependent on these annotated data sets (Torralba & Efros, 2011). Studying social bias within the specific context of age bias and sentiment models brings two major advantages. First, specificity enables me to measure a distinct form of social bias and second, it throws into high relief the implications of age bias for distinct applications.

The primary questions motivating this study are whether age bias manifests in the outputs of language models and, if so, what this bias looks like across commonly-used sentiment analysis models in real-world applications. My analysis focuses both on the treatment by sentiment analysis methods of words that explicitly encode age (e.g., "old" or "young") as well as words that are implicitly encode age (as determined through word embeddings). Given that context is deeply tied to algorithmic bias, researchers have called for technologies to be studied in the contexts in which they operate (Crawford, 2016). I evaluate the impact of these techniques on a text-based corpus of discussions of aging to observe how age bias may manifest in this naturalistic context.

### 3.2.1. Algorithmic Bias in Text Processing

Natural language processing techniques, including sentiment analysis, are a primary point for inquiry among both social scientists and machine learning researchers (Budanitsky & Hirst, 2006; Davidov et al., 2010; Huang et al., 2012; Maas et al., 2011). Sentiment analysis is often used to measure opinions in product reviews or financial markets (Hu & Liu, 2004), which can then inform and drive branding decisions, political campaign strategies, and automated financial trading systems (Feldman, 2013). There are a wide variety of models used for sentiment analysis (or opinion mining), which can be understood as the "computational treatment of opinion, sentiment, and subjectivity in text" (Pang & Lee, 2008). Many sentiment models are lexicon-based, which involves using pre-determined sentiment values of component words and phrases within a document to calculate a sentiment value for the whole (e.g. (Taboada et al., 2011)). Lexicon-based algorithms typically feature a lexicon with words pre-scored for positivity and negativity such

that, when a sentence input is provided, the values of component words are simply summed to produce an overall sentiment score. Another common approach (corpus-based) is to employ predictive classifying techniques, such as supervised learning, to train a machine learning algorithm. Other techniques are hybrids, using some combination of lexicon-based and predictive techniques (Socher et al., 2013; Wilson et al., 2005).

Social bias can arise from a variety of sources in the application of language models. Some work has focused on word embeddings (Bolukbasi, Chang, Zou, et al., 2016; Caliskan et al., 2017), which are multidimensional maps of words and phrases represented by vectors of real numbers. Word embeddings use a variety of text analytics, such as word co-occurrence, to capture the semantic relatedness of terms within a data set and map them in relation to each other. In a word embedding, words located near each other tend to be similar in meaning or how they are used within a data set. Social bias also emerges in algorithmic decision-making, which is often opaque to end users of a given technology (Boyd et al., 2014). Instances of social bias in text processing include the auto-complete function of search engines (Baker & Potts, 2013), advertisements based on search terms (Sweeney, 2013), and image search results (Kay et al., 2015), which can propagate harmful racial and gender stereotypes (Noble, 2018).

To investigate whether age-related bias might be present in sentiment analysis methods, and to understand how various characteristics of sentiment methods influence this form of bias, I study several lexicon-based and corpus-based tools, the type of data they were validated against, as well as word embedding models upon which many algorithmic tools are built. I view each of these features as an area where bias can be introduced, amplified, or potentially diminished.

### 3.2.2. Research Approach

Throughout Study 1, I use a three-phased approach to understand whether and how different sentiment analysis methods produce bias in their outputs with respect to age. First, I examine the extent to which popular sentiment analysis tools exhibit age bias around explicit encodings of age contained within actual sentences sampled from a realistic research context – a community of older adult bloggers. In other words, I sampled cases when age is clearly and unambiguously mentioned in written language (e.g., “It’s starting to be a trend to lay off older workers”). Next, I explore the extent to which the same sentiment analysis methods may demonstrate age bias derived from more implicit encodings of age. To achieve this, I make use of commonly-used word embeddings to produce a set of “older” and “younger” variants of common English adjectives (e.g., for the word “unique”, the “older” analog is “distinctive” while the “younger” analog is “innovative”), and then compare outputs from sentiment analysis tools in a similar fashion to that used in

| Model | Type | Data Source |
|---|---|---|
| *AFINN* | Lexicon | Social Media |
| *EmoLex* | Lexicon | Other |
| *HappinessIndex* | Lexicon | Other |
| *NRC Hashtag* | Lexicon | Social Media |
| *Opinion Lexicon* | Lexicon | Other |
| *PANAS* | Lexicon | Social Media |
| *Sasa* | Classifier | Social Media |
| *Sentiment140* | Classifier | Social Media |
| *SentiStrength* | Hybrid | Other |
| *Sentiwordnet* | Hybrid | Other |
| *SOCAL* | Lexicon | Other |
| *Stanford* | Hybrid | Other |
| *Umigon* | Lexicon | Social Media |
| *VADER* | Lexicon | Other |

Table 3.1. The 15 sentiment analysis methods examined, and their corresponding type and validation data used when building the model. Validation data that is not social-media-based is predominantly based on movie or product reviews or news corpora.

the first analysis. Finally, I train a custom sentiment model by selectively sampling an existing Twitter data set in an attempt to address bias in training data. Taken together, my approach provides a robust assessment of age-related bias in sentiment analysis models across a number of tools, with a variety of sentence types and forms, and begins to assess methods for addressing observed bias.

## 3.3. Phase 1: Explicit Encoding of Age

The goal of my first phase of analysis is to determine whether sentiment analysis tools treat explicit indications of age (e.g., "old" and "young") differently.

### 3.3.1. Method

I perform my analysis using fifteen popular sentiment analysis tools used in practice (Table 3.1 describes the tools and associated annotations that I use). There is no single tool or standard used in sentiment analysis. I explore a variety of sentiment analysis tools to minimize the likelihood of reporting idiosyncratic findings from a single tool, and so that I can compare the effect of common implementation techniques (e.g., lexicon-based vs. supervised learning algorithms) on output bias. I use 15 of the 20 sentiment analysis models implemented in SentiBench (Ribeiro et al., 2016) that span a variety of computational techniques, domains, and levels of complexity. I exclude the remaining five models due to a

lack of variance in output scores and because one model only accepts emoticons as input. In line with how sentiment analysis models are often used, the tools are standardized to produce one of three sentiment outputs: negative (-1), neutral (0), or positive (+1). In my analyses, I also code each sentiment analysis tool according to its computational method (unsupervised, lexicon-based approach vs. supervised, corpus-based approach) and the type of training and validation data underlying the model (social media vs. other sources). Because my method is an initial probe for sources of bias, I do not code for all model characteristics; however, my work sets the stage for more detailed analyses to consider additional facets of model architecture and data processing.

### 3.3.2. Statistical Methods

I test the sentiment tools for age-related bias by examining the sentiment output scores using multinomial log-linear regressions (via the R package nnet (Venables & Ripley, 2002)). I build two types of multinomial log-linear regressions: 1) a single full model for each phase of analysis that includes the data from all of the sentiment analysis tools in order to test for the presence of age-related bias across the models (Table 3.2), and, 2) individual models for each sentiment analysis tool (15 in total) in order to assess how age-related bias may vary across each specific tool (Table 3.3).

My dependent variable is sentiment output (nominal: “negative”, “neutral”, “positive”) and my primary independent variable of interest is the relative age of the adjective in the sentence (”old” vs. “young”). I also examine the regression coefficients to assess how they vary across the different sentiment analysis tools according to the type of sentiment tool used (lexicon-based vs. corpus-based), and each model’s validation data (social media vs. other data).

Regarding regression result interpretation, the exponentiated coefficients represent relative risk (the statistical models are log-linear and produce multinomial logit coefficients)– or the change in the model prediction (i.e. holding all other variables constant, the relative change in likelihood for a sentence to be classified as “positive” or “negative” sentiment rather than default “neutral”). Neutral sentiment is the reference category used by the regressions for all but one of the tools. Exponentiated coefficient values greater than one indicate that the regression model’s sentiment output is more likely and a “neutral” sentiment prediction is less likely, and exponentiated coefficient values less than one indicate that the regression model’s sentiment is less likely and a “neutral” sentiment prediction is more likely.

### 3.3.3. Context of Study and Testing Data

For this analysis, I focus on input sentences that discuss age and aging. I sourced sentences for the analysis by scraping 4,151 blog posts from a prominent “elderblogger” community (Lazar et al., 2017) as well as 64,283 comments on

each post created between 2004 and 2016. To generate the test data set, each researcher then independently, randomly sampled posts and comments containing sentences with the word "old". Of these posts, I extracted 162 unique sentences. Because I am particularly interested in the use of the term "old" to describe people and aging, I exclude sentences using "old" to modify nouns other than people (e.g., "old things", "old movie") and as a general descriptor of age (e.g., "the 32-year-old"). In the sample I exclude sentences that contain the word "young" or other youth-related terms in order to focus on the ways in which *older* age is discussed. I also exclude complex sentences with embedded clauses or unusual grammar or structure. Example sentences include, "I live in a culture that deliberately hides and ignores older folks." and "Old age is worth waiting for." Although the term "old" appears 86,145 times across my corpus, random sampling and my exclusion process resulted in 121 sentences from my initial sample.

For comparison, I next create a near-identical test set focused on terms referencing young age. In each of the original 121 sentences, I replace the term "old" (as well as "older" and "oldest") with the term "young" (as well as "younger" and "youngest") to provide a comparative data set. In the end I have a test set with 242 sentences total– 121 "old" sentences, and 121 "young" sentences [2]. My goal in doing this is to understand if sentiment analysis tools provide equivalent sentiment measures if the content from this blog were to describe younger people and youth instead of old age.

By using a standardized set of sentences and varying only the age-related terms, I am able to attribute any observed changes in sentiment outputs to the particular aged words I vary. Example sentences include, "It also upsets me when I realize that society expects this from <old/young> people."; "But it is not <old/young> folks who should be ashamed and embarrassed; it is the culture at large." I test all 242 sentences by running each through the 15 sentiment analysis tools. This produced 3,630 sentiment analysis outputs.

### 3.3.4. Results

My findings for this phase of analysis are threefold. First, the results of the regression (see details in Table 3.2) revealed that across all of the sentiment analysis tools, sentences containing young adjectives (*"Young" Adj.*) were 66% more likely to be scored positively than the same sentences containing old adjectives, when controlling for other sentential content. If there were no age bias in sentiment outputs, I would expect to see an equal likelihood for a sentence to be scored as "positive", "neutral", or "negative", regardless of whether it contained "old" or "young".

---

[2] Appendix 8.5 shows the sentences in this test set as well as the test set used in later analyses of both this study and the following study

| Output | "Young" Adj. | | Corpus-Based Model | | Social Media Source | | "Yng" X Corpus-Based | | "Yng" X SM Source | | Intercept | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *e^(coef)* | *95%CI* | *e^(coef)* | *95%CI* | *e^(coef)* | *95%CI* | *e^(coef)* | 95%CI | *e^(coef)* | *95%CI* | *e^(coef)* | *95%CI* |
| *Positive* | **1.66**** | [1.3-2.11] | **2.56**** | [1.99-3.29] | **0.51**** | [0.39-0.65] | **0.59**** | [0.42-0.85] | 0.84 | [0.60-1.18] | **0.76**** | [0.63-0.90] |
| *Negative* | 0.88 | [0.68-1.15] | **2.73**** | [2.17-3.45] | 1.14 | [0.91-1.42] | 1.21 | [0.87-1.70] | 0.98 | [0.71-1.35] | **0.70**** | [0.59-0.84] |

Table 3.2. Regression results for explicit age analysis. The results indicate the likelihood of a "positive" or "negative" outcome (rather than "neutral") given the variable input at the top of each column. The models include data from all sentiment analysis tools and are multinomial log-linear regressions, resulting in a model for positive sentiment and a model for negative sentiment. The reference categories are: neutral sentiment, "old" adjectives (i.e., "old" or "older"), lexicon-based approaches, and non-social-media validation data. Exponentiated coefficients (i.e., e^coef) provide relative risk (e.g., the sentiment analysis models were 1.66 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "older" adjective to a "younger" adjective"). Note: *$p$<0.05; **$p$<0.01

Second, examining the type of sentiment analysis tools, supervised learning-based tools (Corpus-based, as opposed to lexicon-based) were more likely to indicate either positive or negative sentiment (rather than neutral) compared with unsupervised, lexicon-based tools, indicating a polarizing effect. Because supervised learning-based tools had a polarizing effect on the likelihood of both "positive" and "negative" outputs, and because the sentiment analysis tools were more likely to indicate "positive" for "young" sentences, these two trends had a disproportionate effect on pushing "young" sentences toward "positive" sentiment. Third, sentiment analysis tools validated against social media data (*Social Media Source*) were overall less likely to rate sentences as "positive" (rather than neutral) compared with tools validated against other forms of data.

These findings may be explained by the fact that machine learning classifiers are necessarily trained using extremely large corpora of natural text produced by people (e.g., social media posts, news articles, movie reviews). Training on naturalistic text allows classifiers the ability to learn subtle biases in human language use. Much of this training data is social media-based (e.g., Twitter) and may contain a wide variety of social and political views.

Analyzing the outputs from all 15 sentiment models reveals a significant interaction between age and the type of sentiment algorithm (lexicon-based vs. corpus-based). The corpus-based method diminished the likelihood of a positive outcome for "young" sentences compared with the lexicon-based method. Considering the individual regression model results provides a more complete picture by partitioning out the results by each sentiment analysis tool (see Table 3.3 for details). These results reveal that 4 (of 15) sentiment methods were significantly more likely to indicate positive sentiment when a sentence contained a young-age-related adjective. Three of these tools were lexicon-based and one was corpus-based. Two tools were less likely to indicate negative sentiment for young-age-related sentence and one was more likely to indicate negative sentiment for a "young" sentence.

Taken together, the results of both the full regression and the individual regressions indicate significant age-related bias with respect to explicit encodings of age, with corpus-based tools presenting a more polarizing effect and those trained on social media data skewing less positively across all sentences. Yet, questions remain around whether similar bias exists for implicit encodings of age, such as through the word embeddings that underlie many of these tools.

## 3.4. Phase 2: Implicit Encoding of Age

Age-related bias may seep into computational approaches in various ways. Because "old" and "young" are simple, explicit references to age, my prior analysis does not fully account for other ways that age is discussed and referenced in language. To build on my first phase of analysis I consider the ways in which age is *implicitly* referenced. As an example of implicit association, Axelson et al. (2010) studied adjectives used in evaluations of medical students, finding that women were more likely to be described as "compassionate" and "sensitive" compared with men exhibiting similar performance. "Compassionate" and "sensitive" are not inherently gendered terms; however, they align with prevailing gender stereotypes about women as emotional that can hinder women's career progression (Brescoll, 2016). For the second phase of this study I analyze whether words implicitly associated with younger and older age are also treated differentially by sentiment algorithms.

### 3.4.1. Method

I repeat my process from phase 1 and again manipulate test sentences by swapping out "old" and "young". However, I now generate the new adjectives to insert into the templates by taking a list of common English adjectives and deriving implicitly "old" and "young" variants of each word. In order to derive words implicitly associated with age, I rely on word embeddings. Word embeddings are multi-dimensional vectors (often 100-300 dimensions) where each vector represents a specific word, and the values for each dimension are calculated based on the context (i.e., surrounding words) within which that word commonly appears. One of the most salient emergent properties of word embeddings is that they have been shown to encode analogies (e.g., "king" – "man" + "woman" = "queen") (Mikolov et al., 2013). Thus, word embeddings can be used to transfer the semantic relationship between two words (e.g., between "man" and "woman") onto a different word (e.g., "king") and provide a reasonable semantic analog (i.e., "queen").

While word embeddings are effective at capturing semantic and syntactic properties of words, they also have been shown to latently encode stereotypes and human biases (e.g., "computer programmer" – "man" + "woman" = "homemaker") (Bolukbasi, Chang, Zou, et al., 2016). I explore this in the context of age and generate "older" and

| | Sentiment Analysis Tool | | Positive | Intercept | Negative | Intercept |
|---|---|---|---|---|---|---|
| Lexicon | AFINN | e^(coef) | 1.000 | 0.956 | 1.000 | 0.733 |
| | | 95% CI | [0.553 - 1.807] | [0.63 - 1.451] | [0.53 - 1.887] | [0.468 - 1.149] |
| | EmoLex | e^(coef) | **3.180*** | 1.119 | **0.368**** | 0.762 |
| | | 95% CI | **[1.732 - 5.839]** | [0.738 - 1.695] | **[0.141 - 0.958]** | [0.481 - 1.208] |
| | Happiness Index | e^(coef) | **3.743**** | **1.972**** | **2.373*** | **0.389**** |
| | | 95% CI | **[1.852 - 7.565]** | **[1.319 - 2.947]** | **[1.458 - 6.435]** | **[0.21 - 0.721]** |
| | NRC Hashtag | e^(coef) | 1294.656 | 7122.400 | 0.001 | 8.871 |
| | | 95% CI | [0 - 6.9E+25] | [0-3.7E+26] | [0 - 5.3E+19] | [0 - 4.0E+27] |
| | Opinion Lex | e^(coef) | 1.000 | **0.600*** | 1.000 | **0.600**** |
| | | 95% CI | [0.544 - 1.84] | **[0.39 - 0.923]** | [0.544 - 1.84] | **[0.39 - 0.923]** |
| | Panas | e^(coef) | N/A | N/A | 1.000 | **0.034**** |
| | | 95% CI | N/A | N/A | [0.244 - 4.093] | **[0.013 - 0.093]** |
| | SOCAL | e^(coef) | **2.586*** | **1.933**** | 1.043 | **1.394*** |
| | | 95% CI | **[1.109 - 6.031]** | **[1.036 - 3.605]** | [0.654 - 1.663] | **[1.001 - 1.941]** |
| | Umigon | e^(coef) | 0.999 | **0.243**** | 1.000 | **0.392*** |
| | | 95% CI | [0.482 - 2.071] | **[0.145 - 0.407]** | [0.545 - 1.836] | **[0.255 - 0.602]** |
| | VADER | e^(coef) | 1.000 | **0.238**** | 1.000 | **0.202**** |
| | | 95% CI | [0.502 - 1.994] | **[0.146 - 0.388]** | [0.479 - 2.09] | **[0.12 - 0.341]** |
| Corpus | Opinion Finder | e^(coef) | 0.985 | **0.323**** | 0.957 | **0.538**** |
| | | 95% CI | [0.491 - 1.975] | **[0.198 - 0.528]** | [0.534 - 1.716] | **[0.357 - 0.813]** |
| | Sasa | e^(coef) | 1.034 | **2.578**** | **0.597*** | **2.113**** |
| | | 95% CI | [0.467 - 2.291] | **[1.519 - 4.376]** | **[0.38 - 0.937]** | **[1.453 - 3.072]** |
| | Sent140 | e^(coef) | 1.307 | **6.501**** | 0.963 | **52.98**** |
| | | 95% CI | [0.162 - 10.56] | **[1.466 - 28.835]** | [0.133 - 6.972] | **[13.07 - 214.72]** |
| | Senti Strength | e^(coef) | 1.000 | **0.634**** | 1.000 | 0.692 |
| | | 95% CI | [0.539 - 1.854] | **[0.41 - 0.982]** | [0.548 - 1.825] | [0.452 - 1.059] |
| | Stanford | e^(coef) | 1.111 | 0.834 | 0.797 | **3.209**** |
| | | 95% CI | [0.496 - 2.486] | [0.46 - 1.51] | [0.421 - 1.51] | **[2.029 - 5.077]** |

Table 3.3. Individual regression results for explicit age analysis. The results from each sentiment analysis method were fit to a multinomial log-linear regression model, resulting in a model for positive sentiment and a model for negative sentiment for each sentiment analysis method. The reference categories for each model are: neutral sentiment and "old" adjectives. Coefficients that are not significant at $p<0.05$ are greyed out. Exponentiated coefficients (i.e. ecoef) provide effect sizes for relative risk (e.g. the EmoLex model was 3.18 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from "old" (or "older" or "oldest") to "young" (or "younger" or "youngest") holding all else constant. Note: *$p<0.05$; **$p<0.01$

The Sentiwordnet model (corpus-based) is not included because it did not classify any sentences "neutral." Instead, I used "negative" as the reference category for regression. The model was 4.121 times more likely to indicate positive for a "young" sentence compared to an "old" sentence ($p<0.01$, 95%CI: [2.390, 7.106]).

| Embedding | Source | Vocabulary |
|---|---|---|
| WG-6B-50D | English Wikipedia 2014 text and Gigaword 5 (7 sources of English-language newswire) | 400K words, uncased |
| WG-6B-100D | | |
| WG-6B-200D | | |
| WG-6B-300D | | |
| CC-42B-300D | Common Crawl of the Internet | 1.9M words, uncased |
| CC-840B-300D | | 2.2M words, cased |
| TW-27B-25D | 2 billion tweets | 1.2M words, uncased |
| TW-27B-50D | | |
| TW-27B-100D | | |
| TW-27B-200D | | |

Table 3.4. Details on the 10 GloVe models. The first part of the name references the source, the second part of the name gives the number of tokens contained in the source (e.g., 6B = 6 billion), and the third part of the name gives the number of dimensions of the word vectors (e.g., 300D = 300-dimensional vectors for each word in the vocab). Further details at https://nlp.stanford.edu/projects/glove/

"younger" analogs of common adjectives. I start with the 500 most common English adjectives (Davies, 2008) and then generate "older" and "younger" analogs for each adjective. For example, I find in one embedding that "stubborn" – "young" + "old" gives "obstinate" while "stubborn" – "old" + "young" gives "courageous". As a control, I also identify the word mapped nearest to the original adjective (e.g., in this case, also "obstinate" for "stubborn"). I do this because, in the absence of age bias, the word embeddings produce the nearest adjective rather than the original input adjective. Using the nearest adjective as the reference category allows my statistical model to account for instances where no age bias is detected. I then substitute these three versions of each adjective into my template sentences (i.e., the control adjective, the "older" adjective, and the "younger" adjective). I test 10 different word embedding models, the common GloVe (Global Vectors for Word Representation) embeddings provided by Pennington et al. (2014). These embeddings differ in the number of dimensions (fewer dimensions encode less information about a word) and text from which they were trained (Wikipedia, a common crawl of the Internet, and Twitter). See Table 3.4 for a description of each embedding. Similar to phase 1, I test word embedding features to probe possible sources of bias.

As in phase 1, I classify each sentence according to each of the 15 sentiment analysis tools. In order to keep the number of sentences and sentiment analysis outputs to a computationally tractable level, I used three researcher-generated sentence templates in this analysis ("The <adj><noun> went to the movies", "The <adj><noun> had a lot of trouble understanding. "The <adjective><noun> wrote an amazing novel"). In addition to varying "young" and "old"

variants of each adjective, I vary the gendered noun being modified (e.g. “man”, “woman”, “person”). I do this based on work by Lazar et al. (2017) in which older adult women note unique experiences rooted in their social identities as older adult women, rather than older adult men or younger women. In the end, my process results in 135,000 sentences in total (3 templates x 500 adjectives x 3 adjective types x 10 word embeddings x 3 nouns); running each through all 15 sentiment analysis tools results in 2,025,000 sentiment analysis outputs.

### 3.4.2. Results

In line with the results from phase 1, which found significant differences in the sentiment of explicit age-related keywords, I also found significant differences in the sentiment of implicitly coded age-related keywords generated through word embeddings. The full regression results indicated that sentences constructed with implicitly “old” adjectives were 0.91 times as likely to be scored positive, compared with the control adjective ($p$<0.01, 95% CI [.899, .921]). Similarly, sentences with implicitly “old” adjectives were 1.03 times more likely to be scored more negatively compared with the control adjective ($p$<0.01, 95%CI [1.017, 1.045]). Sentences with implicitly “young” adjectives were 1.09 times more likely to be scored positive ($p$<0.01, 95% CI [1.075, 1.101]). And sentences with implicitly “young” adjectives were 0.94 times as likely to be scored negatively ($p$<0.01, 95% CI [.926, .952]).

I included all 10 GloVe word embeddings in the full regression, and examined whether there was variation in effects across the different word embeddings. Although I could not definitively isolate which embedding source yielded the most bias (because differing dimensionality across embeddings meant that I could not directly compare all of them), the Wikipedia embeddings demonstrated the least amount of bias, whereas Twitter embeddings led to the greatest bias. Examining the individual regressions (Table 3.5), 9 of 15 models indicate a significantly greater likelihood of positive sentiment in implicitly “young” adjectives as compared to the control adjective (*“Young” Adj.*). Similarly, 12 of 15 models exhibit a significantly lower likelihood of indicating positive sentiment for the “old” adjectives compared to the control (*“Old” Adj.*). In terms of negative sentiment, 11 of 15 models see a significantly lower likelihood of indicating negative sentiment for implicitly “young” adjectives (*“Young” Adj.*), but there were mixed results for the effect of implicitly “old” adjectives on the likelihood of indicating negative sentiment. On the whole, the sentiment of implicitly “young” adjectives generated through the word embeddings were more likely to be rated positively and less likely to be rated negatively compared to implicitly “old” adjectives.

| | Sentiment Analysis Tool | | "Positive" Likelihood | | | "Negative" Likelihood | | |
|---|---|---|---|---|---|---|---|---|
| | | | "Young" Adj. | "Old" Adj. | Intercept | "Young" Adj. | "Old" Adj. | Intercept |
| Lexicon | AFINN | e^(coef) | 1.01 | **0.908**** | **1.347**** | **0.891**** | 0.969 | **1.17**** |
| | | 95% CI | [0.979 - 1.042] | **[0.878 - 0.939]** | **[1.308 - 1.387]** | **[0.862 - 0.921]** | [0.937 - 1.002] | **[1.134 - 1.207]** |
| | EmoLex | e^(coef) | **1.134**** | **0.848**** | **0.739**** | **0.783**** | **0.953*** | **0.112**** |
| | | 95% CI | **[1.103 - 1.166]** | **[0.825 - 0.872]** | **[0.72 - 0.758]** | **[0.737 - 0.832]** | **[0.902 - 1.007]** | **[0.106 - 0.118]** |
| | Happiness Index | e^(coef) | 1.017 | **0.08**** | **7E+07**** | 0.973 | **0.082**** | **3E+07**** |
| | | 95% CI | [0.205 - 5.054] | **[0.025 - 0.26]** | **[3E+07 - 2E+08]** | [0.196 - 4.837] | **[0.025 - 0.266]** | **[1E+07 - 7E+07]** |
| | NRC Hashtag | e^(coef) | **1.058**** | 1.008 | **3.347**** | 0.991 | 1.020 | **2.088**** |
| | | 95% CI | **[1.005 - 1.113]** | [0.958 - 1.008] | **[3.218 - 3.481]** | [0.957 - 1.009] | [0.967 - 1.02] | **[2.004 - 2.176]** |
| | Opinion Lex | e^(coef) | **1.110**** | **0.929**** | **1.033*** | **0.882**** | **1.077**** | **0.968*** |
| | | 95% CI | **[1.076 - 1.145]** | **[0.9 - 0.959]** | **[1.003 - 1.064]** | **[0.853 - 0.912]** | **[1.044 - 1.111]** | **[0.94 - 0.997]** |
| | Panas | e^(coef) | **5.876**** | **0.157**** | **0.004**** | **0.677**** | **0.741**** | **0.01**** |
| | | 95% CI | **[4.994 - 6.914]** | **[0.105 - 0.236]** | **[0.003 - 0.005]** | **[0.586 - 0.783]** | **[0.643 - 0.853]** | **[0.009 - 0.011]** |
| | SOCAL | e^(coef) | **1.124**** | **0.805**** | **1.752**** | **1.041*** | **0.941**** | **1.536**** |
| | | 95% CI | **[1.087 - 1.162]** | **[0.779 - 0.832]** | **[1.698 - 1.808]** | **[1.005 - 1.078]** | **[0.91 - 0.973]** | **[1.489 - 1.585]** |
| | Umigon | e^(coef) | 0.997 | **0.935**** | **1.126**** | **0.950**** | 1.030 | **1.207**** |
| | | 95% CI | [0.964 - 1.031] | **[0.904 - 0.967]** | **[1.093 - 1.16]** | **[0.921 - 0.98]** | [0.998 - 1.063] | **[1.172 - 1.243]** |
| | VADER | e^(coef) | **1.071**** | 0.973 | **0.525**** | **0.626**** | **1.089*** | **0.059**** |
| | | 95% CI | **[1.042 - 1.101]** | [0.947 - 1] | **[0.512 - 0.539]** | **[0.579 - 0.677]** | **[1.017 - 1.166]** | **[0.055 - 0.063]** |
| Corpus | Opinion Finder | e^(coef) | **1.581**** | **0.613**** | **0.026**** | **0.787** | 0.996 | **0.057**** |
| | | 95% CI | **[1.47 - 1.7]** | **[0.56 - 0.671]** | **[0.024 - 0.028]** | **[0.733 - 0.844]** | [0.932 - 1.065] | **[0.054 - 0.06]** |
| | Sasa | e^(coef) | **1.175**** | **0.824**** | **0.22**** | **0.858**** | **0.950**** | **0.61**** |
| | | 95% CI | **[1.132 - 1.22]** | **[0.792 - 0.857]** | **[0.212 - 0.228]** | **[0.832 - 0.885]** | **[0.922 - 0.978]** | **[0.593 - 0.627]** |
| | Sent140 | e^(coef) | 1.293 | 1.287 | **483**** | 1.297 | 1.309 | **237**** |
| | | 95% CI | [0.825 - 2.025] | [0.823 - 2.012] | **[338.7 - 688.7]** | [0.828 - 2.032] | [0.836 - 2.051] | **[166.2 - 337.9]** |
| | Senti Strength | e^(coef) | 1.030 | **0.925**** | **1.207**** | **0.919**** | 1.006 | **1.195**** |
| | | 95% CI | [0.998 - 1.063] | **[0.895 - 0.956]** | **[1.172 - 1.243]** | **[0.889 - 0.95]** | [0.973 - 1.04] | **[1.16 - 1.231]** |
| | Sentiwordnet | e^(coef) | 0.979 | **0.766**** | **3.216**** | **0.861**** | **0.900**** | **2.795**** |
| | | 95% CI | [0.934 - 1.026] | **[0.731 - 0.803]** | **[3.092 - 3.345]** | **[0.821 - 0.902]** | **[0.859 - 0.943]** | **[2.688 - 2.907]** |
| | Stanford | e^(coef) | **1.142**** | **0.884**** | **0.598**** | **0.783**** | **1.788**** | **0.042**** |
| | | 95% CI | **[1.111 - 1.174]** | **[0.86 - 0.909]** | **[0.583 - 0.613]** | **[0.709 - 0.865]** | **[1.666 - 1.919]** | **[0.039 - 0.045]** |

Table 3.5. Individual regression results for the implicit age analysis. The results from each sentiment analysis method were fit to a multinomial log- linear regression. The reference categories for each model are: neutral sentiment, and "control" adjectives. Exponentiated coefficients (i.e. e^coef) provide effect sizes for relative risk (e.g. the top right coefficient -0. the EmoLex model was 1.134 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "control" adjective to an "older" adjective as determined by the word embeddings. Note: *$p$<0.05; **$p$<0.01

## 3.5. Phase 3: Addressing Age Bias via Training Data

Given that the first two phases of my work reveal the existence of age bias in sentiment analysis models, the final phase aims to demonstrate a technique to mitigate that bias. By mitigating bias present in existing tools, researchers might still take advantage of these computational approaches to study topics where attitudes toward age matter. In this third and final phase, I modify the training data set originally used to create the Sentiment140 classifier and train my own custom models with this modified data. This allows us to conduct a more fine-grained analysis of bias within a single model to then trace and remove it.

### 3.5.1. Method

First, I build two custom sentiment classifiers. There are two components to each classifier: the model architecture and the data upon which they are trained. Each of my custom models share the same architecture and only vary in the data that I use to train them. This allows us to directly connect output bias to changes in the training data.

Each of my custom models is a Maximum Entropy bag-of-words classifier. This architecture is widely-used in various text classification problems, including sentiment analysis, and predicts the most likely label (e.g. "positive" or "negative") for a given input using logistic regression. Bag-of-words models convert text inputs to a set of words, disregarding word order and grammar but retaining word frequency. This set of words is used as an input to the model, which then learns how different patterns of words map to the different labels across thousands of inputs. I use the Python SciKit Learn package—a common machine learning package—to create and train my models.

For training data, I required a data set of annotated text that I could manipulate for my custom classifiers. I adopt the training data used by Sentiment140 because it is one of only two publicly-available, annotated training data sets used to train a sentiment model that I tested. The training data was annotated through an automated process wherein tweets were annotated based on an initial set of researcher-annotated tweets as well as the presence of emoticons (Go et al., 2009). This original training data set contains 1.6 million tweets and corresponding labels.

I split the original Sentiment140 training data set into two, exclusive subsets to observe whether I can isolate bias in the training phase of creating the classifier. First, I filter the training data to find tweets that include the terms "young" and "old". This leaves me with a training data set of 13,781 tweets, which I refer to as the *Age-Only* corpus [3]. I use

[3]Although "old" and "young" have several definitions and are not always used to describe humans, my custom classifier does not feature word sense disambiguation. However, I created an additional data set filtered by age- related phrases to isolate uses of "old" and "young" strictly with respect to humans (e.g., "young man", "older people"). This data set was much smaller than the others we produced (1,550 training examples) and produced outputs similar to those of our data set filtered on the terms "old" and "young."

this data set to determine where bias exists. I then reverse this filtering process to create a second data set that excludes these age-related tweets (referred to as the *Age-Removed* corpus). This data set allows me to diagnose the extent to which bias in the *Age-Only* corpus impacts output bias. I also retain the original, unfiltered data set to implement the *Original* classifier. The *Original* classifier stands as a control against which I can compare my manipulations.

Similar to the first phase of my analysis, I run each of my custom-trained models on a test set of sentences sourced from the posts and comments of a blog written by older adults. Starting from the original 121 sentences I used in phase 1, I randomly sampled and filtered additional sentences to obtain 169 sentences containing the term "old", I duplicate the sentences and replace the term "old" with "young" to double the set to 338 sentences, which are then used to test the custom classifiers for the presence of bias (i.e. difference between output probabilities for "old" and "young" sentences). I increase the sample size to provide greater sensitivity and to help illuminate whether my filtering approaches could be effective. Departing from my phase 1, I analyze the outputs from each of the custom-trained models using a paired t-test to determine the extent of bias that results from training on each of the different corpora. Specifically, given an input sentence, my classifiers provide a probability (otherwise known as model confidence) that the correct classification should be "positive". Unlike previous phases, I use this continuous confidence, rather than a categorical output of 'positive' or 'negative' in my statistical model. For example, my *Original* model (i.e. trained on the unfiltered training data) classifies the sentence, "As much as I work on acceptance of getting old, I don't like it!" as "negative" with a 0.9509 probability (i.e. 95% confidence) and just 0.0491 probability (i.e. 5% confidence) in the alternative outcome ("positive"). If there were no bias (i.e. if the classifier treated "old" and "young" as equivalent in sentiment), I would expect an equal number of positive outcomes for "old" and "young" sentences. Although the Sentiment 140 model did not exhibit statistically significant bias in the phase 1 analysis, the model estimates trended in the expected direction. Additionally, in phase 3 I use a larger data set (169 sentences vs. 121 sentences in phase 1) and use a continuous probability rather than categorical output, which provide a more sensitive measure for bias. Notably, while the models I test in phases 1 and 2 also include "neutral" as a class, my custom implementation only differentiates explicitly between "positive" and "negative." This implementation choice is driven by the fact that, although the Sentiment140 model produces a "neutral" outcome, the training data set made publicly available only features "positive" and "negative" annotations.

By isolating the age-related tweets in my different training corpora, I can determine the source of the output bias and assess whether manipulating examples of "old" and "young" can effectively prevent my custom classifier from

| **Train Data** | *Original* | *Age-Only* | *Age-Removed* |
|---|---|---|---|
| **Increase in likelihood for a "young" sentence to be classified as "positive"** | +13.61% | +24.26 | +1.18% |

Table 3.6. The increase in likelihood that a "young" sentence will be classified as "positive" compared to its "old" counterpart. Training the model on the full, original dataset, a "young" sentence was 13.26% more likely to be "positive" compared to its "old" counterpart. There were 169 "old" and "young" sentence pairs.

| | *Original* | *Age-Only* | *Age-Removed* |
|---|---|---|---|
| **Mean Confidence "young"-"pos"** | 0.5867 | 0.5161 | 0.5671 |
| **Mean Confidence "old"-"pos"** | 0.5196 | 0.4492 | 0.5608 |
| **Mean Di□erence** | 0.0671 | 0.0669 | 0.0063 |
| *[95%CI]* | *[0.023,0.111]* | *[0.023,0.111]* | *[-0.038,0.050]* |
| **p-value** | $p<0.0027$* | $p<0.0028$* | $p<0.7796$ |

Table 3.7. Paired t-test results for the custom-trained classifiers. A likelihood above .50 produces a classification of "positive."

exhibiting age-related patterns of bias possibly rooted in these training examples. If I observe the greatest bias in the *Age-Only* and *Original* corpora, this would indicate that the output bias is embedded in the labels of these age-related tweets. If I observe the greatest bias in the *Age-Removed* corpus; however, this would indicate that the output bias results from a dearth of training examples related to age and aging. Finally, if there is no significant difference in bias between the *Original* and *Age-Removed* corpora, this would indicate that the output bias largely derives from other, less contextually relevant tweets in the original data set. Worth noting is that my approach addresses a reduction in explicit age-related bias, rather than implicit bias, which may manifest as coded language or stereotyping.

### 3.5.2. Results

Table 3.6 shows the increase in likelihood for a sentence to be classified as "positive" when "old" is replaced with "young". Overall, I find the greatest output bias (as indicated in mean difference in Table 3.7 in classifiers trained on the *Age-Only* and *Original* corpora (both of which contain tweets with "old" and "young") and no significant bias in the *Age-Removed* corpora. This indicates that the age bias does indeed originate from bias in the annotations of the age-related tweets and can be remedied by removing these training examples.

The custom classifier trained on the *Original* data set produced significant age bias with respect to the terms "old" and "young" ($p$<.0027) where sentences containing the terms "old", "older", or "oldest" were more likely to be classified as "negative". This result is in line with that of my phase 1 aggregated analysis. The custom classifier trained on the *Age-Only* corpus also produced significant bias ($p$<.0028). This classifier produced outputs with lower confidence in a "positive" classification compared to the custom classifier trained on the full Sentiment140 data set. This result indicates that the age-related tweets in the training data were more negative than the overall corpus.

The custom classifier trained on the *Age-Removed* corpus did not show significant bias ($p$=.7796). The reduction in bias compared to the classifier trained on the original data set and the classifier trained on age-related tweets, was statistically significant ($p$<.0008). Notably, the mean gap in likelihood for an "old" vs. "young" sentence to be classified as positive was an order of magnitude lower compared with the other two classifiers (0.0063 vs. 0.0671 and 0.0669). Although the *Age-Removed* model had a slightly higher probability of classifying "young" sentences as positive as compared to "old" sentences as positive, this stemmed from only classifying two (out of 169) sentences containing the term "old" differently than their "young" counterparts. In both of these instances, the classification was negative for the "old" sentence and positive for the "young" sentence.

### 3.6. Strategies for Addressing Bias

This inquiry into age-related bias in sentiment analysis contributes, first and foremost, a systematic analysis of age-related bias in a large number of popular sentiment analysis tools and word embeddings. By carefully and systematically modifying sentence inputs, I found significant age bias in algorithmic output, despite intentionally crafting near-identical test sentences. I also provided a nuanced look at how the technical characteristics of various sentiment models impact bias in outcomes – finding particularly that tools validated against social media data exhibit increased bias. Finally, the overall arc of these analyses stand as a case study of mitigating bias in training data where, with a relatively straightforward approach, I successfully reduced age bias by an order of magnitude. The results of this work beget critical reflection on the use of language models for analyses involving older adults, but also those incorporating social data–especially from social movements and underrepresented populations.

These findings have implications for how researchers interpret sentiment analysis results, the strategies we use to understand and mitigate bias, and the challenges of using these techniques to study underrepresented populations and online social movements. The analysis of my custom trained, maximum entropy models in phase 3 highlighted that I could reduce bias in my chosen research context by re-sampling training data from a larger data set. This relatively

simple change to the training data reduced the age-related bias to statistically insignificant levels and highlights one way in which researchers can isolate where bias emerges in data. It also demonstrates how researchers can begin to account for bias in data sets, as well as how they might adapt available data sets to their particular research context. However, my approach may not work similarly for other types of machine learning models such as those built on recurrent neural networks, which are sensitive to word order and syntax. It also does not address subtler instances of bias, such as the association of broader topics with gender (e.g., women and relationship- and family-related topics) (Wagner et al., 2016). Although I identify implicit age bias in phase 2, mitigating this bias from a data set would likely necessitate techniques more sophisticated than simply removing examples from training data.

My approach is particularly relevant with regard to studying underrepresented populations. When data pertaining to a particular population is sparse or difficult to obtain, adapting a large, existing, annotated data set may be more feasible than collecting sufficient data and annotating it to train a model. Many other approaches also point to underlying bias in the ways certain algorithms operate and generate output. While some researchers consider quantitative approaches to artificially remove bias from a data set (Bolukbasi, Chang, Zou, et al., 2016), such an approach would be difficult to employ across all instances of social bias and neglects the fact that social bias rarely exists along a single dimension, as was indicated by older adult women in Lazar et al. (2017) (i.e., the notion of intersectionality (Crenshaw, 1990)). While data sets can be tailored by sampling from certain communities, the complexity of language makes it virtually impossible to identify and isolate social biases along all dimensions.

Given this complexity, contextualizing how researchers apply, interpret, and report algorithmic outputs is an important step toward avoiding conclusions that a given algorithmic output is universal ground truth or free of social bias. Instead, researchers should view the outputs of a sentiment classifier as an approximation of the subjective opinion of individuals represented in the training data. In the context of my study, this means that, for the classifiers trained on Twitter data, sentiment outputs are a determination of how that particular sample of Twitter users would interpret some input text rather than approximating how the socially underrepresented group would interpret the text. One explanation for why the corpus-based models exhibited age-related bias is because the underlying data sets were also drawn from a predominantly younger demographic present on social media sites (Smith, 2014), and thereby encoding the values embedded in the ways they discuss age.

Regardless, researchers and organizations creating machine learning models can consider adding context to algorithmic outputs by describing the data used in training and the population who generated it. Currently, the origins

of data sets are not always explicit or available (e.g., IBM's Alchemy, Microsoft's Cognitive Services), and the models and data sets may not be modifiable, as was the case for the majority of sentiment analysis models I studied. In fact, only one of the corpus-based models I tested in phase 1 features both publicly available training data and a model that others can re-train on custom data sets. For this reason, it is particularly critical to rethink "off-the-shelf" use of these tools – that is, the use of sentiment analysis models that have not been tailored to the particular context of use. Of course, even in the rare instance that a sentiment model can be modified for a specific context, getting the necessary training data and using it to train a model requires sufficient technical expertise.

### 3.6.1. Probing the Annotation Process More Deeply

While this study provides a first step toward understanding how the technical characteristics of sentiment algorithms affect bias and identifies one technique for reducing bias, there remain questions regarding the provenance of training data used to create the sentiment models. In addition, I did not specifically probe accuracy in this work. That is to say, while I found that sentiment models had a tendency to treat older age more negatively than younger age, I have no reference to assess the accuracy of these outcomes. I pick up this concern in study 2 and specifically look to older adults as a reference against which I can test accuracy. This work involves additional data collection to better understand data annotation. The majority of sentiment models I tested are built on data that is not publicly-available and nothing is known about the annotation population or their attitudes toward aging or the content they annotated. For a higher fidelity look at data collection and annotation, I next shift my investigative lens to data collection, the annotation population, and the performance of models built on their data.

CHAPTER 4

# Bringing Stakeholders into Data

Returning to my opening example in the introduction of this dissertation, if the United States Transportation Safety Administration were to redesign airport body scanners to more inclusively and accurately detect banned foreign objects on trans* and GNC passengers, it is likely that a targeted testing effort with trans* and GNC individuals would be necessary. Representing less than 1% of the United States populations, trans* and GNC individuals are unlikely to be represented in sufficient numbers by generalized data collection or user testing. Rather, a testing scenario must specifically be constructed to assess whether the technology serves them properly. In the same vein, algorithmic systems should be evaluated against underserved stakeholders with targeted testing. Algorithmic systems intended to learn from data representing general human behavior may fail to accurately encode or measure features of underrepresented populations. Targeted analyses of algorithm performance are needed to illuminate whom and which contexts these tools serve best.

Despite this need, many algorithmic tools feature documentation that does not explicitly indicate for whom or for which contexts they have been optimized (Mitchell et al., 2019). Technologies purported to work for broad or general contexts, such as sentiment analysis, behoove investigations of their potential impacts on underrepresented or otherwise marginalized populations. As scholars in Critical Algorithm Studies have and continue to discuss, designers of technologies used at scale too often overlook potential impacts for already-marginalized populations (Ananny, 2011; Baker & Potts, 2013; L. Introna & Wood, 2004). Underrepresented populations often have characteristics, behaviors, or features that, by virtue of being underrepresented, can differ greatly from those of other groups. This has two noteworthy implications. The first is that data describing underrepresented populations may be difficult to obtain or may be altogether absent from the data science pipeline. The second is that, even if data from underrepresented populations is captured in general data collection efforts, there is a possibility that the data points are too few in the data set to be meaningful.

To build off of Study 1 and address questions of representation in data, I focus on involving members from underrepresented populations in the creation and evaluation of sentiment models. Picking up from Sen et al. (2015)'s

assertion that algorithms should be evaluated based on how well they work for a given community, I continue my investigation of age-related bias by soliciting input from older adults. In this work I focus on assembling both testing and training data so that I can trace how learning algorithms encode social bias rooted in underlying data. In the first of my analyses, I use stakeholder input as testing data to assess model accuracy and better understand the effects of age bias in outputs. Next, I use stakeholder input in *training* data to assess if supplementing an existing training data set with stakeholder input might improve model bias and accuracy. In doing so, I respond to the call of D'Ignazio and Klein (2020) to involve underrepresented communities in the data science pipeline– both as a matter of data representation as well as a matter of giving voice and shifting power to these communities. In my investigation of age bias in sentiment analysis, comparing algorithmic outputs to test data solicited from older adults allows me to take a closer look at discrepancies between how sentiment models rate age-related content and how older adults rate the same content. In this way, I frame the data provided by older adults as a snapshot of their expertise and experience with older adulthood.

For this line of inquiry, I draw inspiration from participatory design. While I do not port specific methods in participatory design in this work, I adopt its emphasis on incorporating underrepresented voices and viewpoints into design processes (DiSalvo et al., 2012; Juarez & Brown, 2008). Participatory methods, in general, can be used in a variety of domains to involve end-users and stakeholders in the design and evaluation of products, services, or technologies. These methodologies encompass a range of activities and design exercises that emphasize the voice of stakeholders. Much like Value Sensitive Design, participatory design scholarship, in particular, acknowledges the "inevitable presence of values in the system development process" (Schuler & Namioka, 1993). Participatory design serves as a natural complement to Value Sensitive Design by explicitly highlighting the interplay between design processes and the adoption of stakeholder values in design outcomes. An emphasis on process is of particular import in Study 2 as I shift from evaluating existing algorithmic models to understanding and experimenting with data collection and annotation.

Participatory methods are often employed in the context of digital services and civic technologies as an avenue to underscore the voices and experiences of community stakeholders alongside those of public servants and city project leaders. In his book Designing Publics, Le Dantec (2016) frames participatory approaches as a site for collective action that, *"can be used to draw people together to contend with or resist shared social issues —to participate in the improvement of both their individual conditions and the conditions of their community"*. Invoking this guiding

principle, I frame efforts to engage the perspectives of underrepresented communities as efforts to give them voice in data work. Specifically, I explore approaches to and the implications of representing older adults in sentiment model data. Turning to older adult stakeholders to provide data for model training and evaluation is one way of leveraging their range of perspectives and first-hand experience with older age and aging. My aim in involving older adults is primarily to address their lack of data representation. Older adults are underrepresented on most social media sites from which many training data sets are scraped (Duggan & Brenner, 2013), as well as underrepresented among the crowd working populations that do much of the annotation work featured in these data sets (Posch et al., 2018; Ross et al., 2009). Moreover, the increase in median age of the global population (for Disease Control, (CDC, et al., 2003) means that older adults' perspectives are increasingly salient and important to capture. Building technological systems from data sets that do not substantially represent older adults' values and perspectives is problematic if older adults are expected to be subjects or stakeholders of analyses.

Outside of work specifically related to data-driven systems, researchers in the HCI community have regularly employed approaches, such as the Delphi Method, to leverage the insights of research subjects to improve data analysis (Baumer et al., 2017). In line with Le Dantec (2016) and Woodruff et al. (2018)'s views on incorporating underrepresented voices in research and design, Baumer et al. (2017) state, "The people whom we are studying...may have different interests or perspectives that enable them to corroborate, challenge, deny, or propose alternatives to our interpretations". They highlight an important point that data are not inherently objective or easily interpretable. That is, "the same data can mean different things to different people." In the context of machine learning data sets, soliciting input directly from a population of concern may help researchers capture their direct understanding of some target concept. This stands in contrast to inferring a universal understanding from other groups that may hold incompatible or problematic beliefs. In addition to boosting data representation, involving stakeholders directly captures attitudes toward age and the aging experience that researchers may not be privy to a priori.

### 4.1. Study 2A: Older Adult Input in Evaluation

In the next study I focus on model source data and evaluation (see Figure 4.1). More specifically, I consider testing and training data annotations and the role that older adults can play in shaping model performance.

As a first step toward incorporating input from older adults on age-related content, I look to assembling a test set from older adults' annotations. Test sets are used to measure the accuracy of algorithmic models and are used as ground truth for the outputs models should produce. Notably, in my assessment of age-related bias in sentiment models

I did not conduct any analysis of potential differences in the *accuracy* of outputs of models with respect to age-related bias. That is, my analysis showed that similar sentences were rated more negatively when containing "old" rather than "young", but I made no assessment about which of these outcomes is more accurate. That is, I did not assess whether removing bias improved or deflated accuracy.

In machine learning, accuracy is measured by comparing sentiment outputs for a test set of sentences and comparing those outputs to ground truth annotations representing the correct answers. In this study, I generate a test set of age-related sentences so that I can determine whether sentences rated as "positive" by a custom model, for example are indeed positive. Typically, accuracy is reported as the percent of model outputs that match the corresponding ground truth annotation. By studying model accuracy I can understand the extent to which different methods of mitigating age-related bias might impact both potential bias and overall model performance in different contexts of use.

Test sets are an important component of validating algorithmic models and are used to determine the "correct" behavior that a model should exhibit to be considered viable. For example, a test set for an image recognition model meant to identify domestic animals might include examples of dogs or cats that a model must correctly label. However, for assessments with subjective or unclear "correct" behavior, such as determining whether an age-related remark constitutes age discrimination, ground truth is more complicated to determine. Once again invoking Sen et al. (2015), I select ground truth to evaluate sentiment models' fitness for a specific community– older adults. Older adults stand to be the most negatively impacted by age bias in sentiment models and, for age-related content. In addition, older adults possess nuanced experience with the ways in which older age is characterized– appropriately and inappropriately– by others. Therefore, for age-related content, I defer to older adults as the most relevant experts in interpreting the emotional valence of sentences referencing older age.

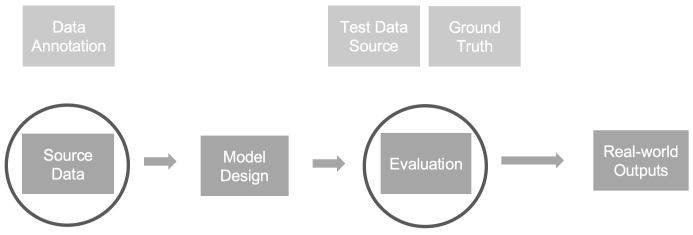


Figure 4.1. Study 2 focuses on real-world applications of computational tools– specifically, real data inputs for analysis and their corresponding outputs.

| Test Data Set | Data Source | Ground Truth |
|---|---|---|
| Age-Related | 296 sentences from older adult bloggers | Panel survey |
| Non Age-Related | 358 sentences from Sentiment140 test set | Sentiment 140 |

Table 4.1. Test data set descriptions.

Allowing the subjects of analysis to contribute ground truth is important for evaluating whether a model aligns with the values and definitions embedded in data from that community. As Woodruff et al. (2018) illustrate, members of underrepresented groups have strong views about how data-driven systems represent their social identities and these views impact trust in these systems, particularly with respect to algorithmic fairness. In their work, the authors make a call to researchers and designers to collaborate with community groups and advocates in designing systems. In particular, they call out community members' experience "considering societal-scale consequences and representing their constituencies on a range of issues." A subtle, yet immensely important, implication in this call to action is an acknowledgement that researchers and designers, on their own, have limited ability to develop robust understandings of technology's broader impacts on underrepresented social groups. At that, attending to differences among older adults, for example, can elucidate differential needs and impacts within an already underserved social group. Technologies used to represent and characterize underrepresented groups, including analytical tools such as sentiment analysis, can benefit from the input and expertise of the very individuals they seek to describe.

### 4.1.1. Test Data

I generate two test sets, shown in Table 4.1, to test the accuracy of my previous custom models. In this analysis, I run two tests of model accuracy– one on age-related content and one on randomly sampled content. As I optimize my custom model for age-related contexts, I also observe changes in performance on non age-related content to understand how my manipulations impact general accuracy. The *Non Age-Related* test set consists of sentences and annotations taken from the Sentiment140 test set. I use all sentences in the publicly-available test set that do not contain the words "old" or "young". Since I was able to remove age-related bias through the *Age-Removed* model, of interest is whether it will perform with different accuracy compared to the *Original* model on each of the test sets. Typically train and test data used in the development of predictive models are derived from the same source. This is done to ensure consistency between the patterns that learning algorithms detect in data and the patterns they will encounter when tested. However, in practice, data analyzed by predictive models can differ substantially from those found in training

data sets. Conducting this analysis allows me to explore whether my method of mitigating age-related bias has any impact on the performance accuracy of the custom model.

The *Age-Related* test set consists of age-related sentences sampled from blog posts authored by older adults and annotated by older adults recruited through a panel survey. This test set seeks to represent older adults both in the test examples as well as in the test annotations. For the age-related analyses in the entire arc of this work, I rely on data and annotations contributed by older adults through a panel survey administered by Qualtrics. Typically, annotation data are not explicitly acquired from specific stakeholders or underrepresented populations. However, I solicit input from a sample of older adults representative of the United States by gender, race, and geographic location (i.e., Northeast, Midwest, West, and South. See Tables 4.2, 4.3, 4.4, and 4.5 for participant details). My method leverages a more representative population of older adults than would be possible using a common approach, such as soliciting annotations through a crowd work platform. In selecting a representative sample, my goal is to capture a wide range of older adult attitudes. I focus on adults aged 50 and older based on prior work I conducted with colleagues that documents age 50 as a common age when age-related workplace discrimination begins in the United States and the United Kingdom (Lazar et al., 2017).

I deploy a single panel survey to obtain annotations for both testing and training data; however, in this initial analysis I focus on test data. I obtain between 4 and 10 annotations per data point using a 5-point Likert scale response

| **Age** | *50-59* | *60-69* | *70-79* | *80-89* | *90-99* | *100+* |
|---|---|---|---|---|---|---|
| **Count** | 542 | 603 | 295 | 40 | 2 | 1 |

Table 4.2. The break down of annotator age.

| **Gender** | *Female* | *Male* | *Nonbinary* |
|---|---|---|---|
| **Count** | 744 | 738 | 1 |

Table 4.3. The break down of annotator gender.

| **Region** | *West* | *Midwest* | *South* | *Northeast* |
|---|---|---|---|---|
| **Count** | 546 | 316 | 546 | 281 |

Table 4.4. The break down of annotator region of residence.

| **Race/ Ethnicity** | *White* | *Black* | *Asian* | *Middle Eastern* | *American Indian/ Alaska Native* | *Native Hawaiian or Pacific Islander* | *Other* | *Hispanic/ Latino (of any race)* |
|---|---|---|---|---|---|---|---|---|
| **Count** | 1,125 | 207 | 96 | 2 | 18 | 4 | 31 | 245 |

Table 4.5. The break down of annotator race and ethnicity. In total there were 1,483 annotators.

(*Very negative* to *Very positive*). I obtained between 4 and 6 annotations for all of the test data points and 99.7% of the training data points. The remaining 0.3% of data points consisted of unremoved duplicates in the original Sentiment140 training data set contained duplicate data points. These 46 items received twice the number of annotations (between 8 and 12) once the duplicates were recognized and the annotation items were aggregated.

In order to choose a single ground truth annotation for each test example, I coded each ordinal response on a scale from -2 to 2 and averaged the score. Each example with an average above 0 was recoded to "positive" and each example with an average below 0 was recoded to "negative". Because the *Original* and *Age-Removed* models produce a binary output (i.e., no "neutral" output), any examples with an average of 0 were removed from the test set. Of the 338 examples in the full *Age-Related* test set, 42 "neutral" examples were removed, leaving 296. Although the Sentiment140 training data set only includes "positive" and "negative" examples, the *test* set includes 139 "neutral" examples. After removing these examples, the final *Non Age-Related* test set contained 358 examples.

### 4.1.2. Models & Architecture

Table 4.6 shows the models I test for accuracy. The models are those that I built and developed in phase 3. As previously described in 3.5, both models are maximum entropy bag-of-words classifiers created using SciKit Learn (Pedregosa et al., 2011). Data for the *Original* model comes from the publicly-available and annotated Sentiment140 training data set. The *Age-Removed* model is built from the same data set, but excluding all training examples that include the words "old" or "young", of which there are 13,781.

### 4.1.3. Accuracy Analyses

I assess overall model accuracy using several measurements. I calculate percent accuracy, precision, recall, and F1 metric for the performance of each model (i.e., *Original* and *Age-Removed*) on each test set (i.e., *Age-Related*, *Non-age-related*). That is, each model will have a set of accuracy metrics for the *Age-Related* test set and an additional set of accuracy metrics for the *Non-age-related* test set.

| Model | Training Data Source |
|---|---|
| Original | Sentiment140 training data |
| Age-Removed | Sentiment140 training data (excluding "old" and "young") |

Table 4.6. The sentiment models and training data sets.

Percent accuracy, precision, and recall are calculated by first summing the number of true positive, false positive, true negative, and false negative outcomes for each sentiment category. True and false rates are calculated by taking the outputs of each model and comparing them to the ground truth labels in each test set. Because of its prevalence as an accuracy measure in sentiment analysis model validations, I also calculate the F1 metric, which is the harmonic mean of precision and recall. I also conduct two McNemar chi-squared tests to analyze the paired, categorical outputs– one comparing each model's outputs on the *Non Age-Related* test set derived from Sentiment140 and one comparing each model's outputs on the *Age-Related* test set derived from older adults' annotations. The results of the McNemar tests will reveal whether the model outputs are significantly different on each of the test sets. That is, even if model accuracy is identical, the McNemar test reveals whether there are significantly different patterns between each model with respect to which test examples they rated correctly and incorrectly. 4.2 provides a visual example of what the McNemar test detects.

Phase 3 demonstrated that the *Original* model exhibited age-related bias and that the *Age-Removed* model did not. Therefore, if the *Original* model performs more accurately on the *Age-Related* test set, that would provide evidence that the *Original* model and the test data encode similar perspectives on aging. Similarly, if the *Age-Removed* model performs more accurately on the *Age-Related* test set, that would suggest that the *Age-Removed* model and the test data encode similar perspectives on aging. However, because the *Age-Removed* model exhibits no significant age bias, higher accuracy would also suggest that the test set does not feature strong trends in positive or negative sentiment with respect to age.


Figure 4.2. Shown are hypothetical outputs of Model A and Model B, where red indicates incorrect classifications. The models rated 4 sentences incorrectly each, producing 60% accuracy. However there is no overlap in which examples they classified incorrectly. A significant McNemar test indicates the existence of many unique, non-overlapping errors in the outputs of each model.

Because the *Original* model and the *Non Age-Related* test set are built from the same data source which was used to validate the Sentiment140 model, the *Original* model should perform with high accuracy on this test set. If the *Age-Related* model also performs accurately, this provides evidence that my modification to reduce age bias may not diminish accuracy for other application contexts.

#### 4.1.4. Results

The McNemar test results indicate that the *Original* model and the *Age-Removed* model performed similarly on each of the test sets. That is, there was no significant difference comparing each model's outputs on the *Age-Related* test set nor on the *Non Age-Related* test set.

#### 4.1.5. Accuracy Results

Table 4.7 shows each model's performance on the *Age-Related* test set and Table 4.8 shows the *Non Age-Related* results. The *Original* model produced higher accuracy on the *Age-Related* test set compared with the *Age-Removed* model; however this difference was only marginally significant (67.23% vs. 64.12%, $\chi^2$=2.783, $p$<0.0953). Between the two models, this translates to differing outcomes on 23 of the 296 sentences. The *Original* model also produced a recall score of 0.787 for the *positive* classification and a recall score of 0.568 for the *negative* classification, as well as an F1 metric of 0.70. The *Age-Removed* model produced a *positive* recall score of 0.787, a *negative* recall score of 0.510, as well as an F1 score of 0.677.

One potential explanation for lower accuracy of the *Age-Removed* model on the *Age-Related* test set is that the ground truth labels generated by older adults align with negative associations with older age encoded in the *Original* model outputs. Upon examining model outputs more closely, their disagreements seem to correlate with whether the test sentence references old or young age. In 16 of the 23 disagreements, the *Age-Removed* model rated the sentence

| Model | Age-Related Accuracy | F1 | Pos Precision | Neg Precision | Pos Recall | Neg Recall |
|---|---|---|---|---|---|---|
| *Original* | 67.23% | 0.70 | 0.624 | 0.746 | 0.787 | 0.568 |
| *Age-Removed* | 64.20% | 0.677 | 0.594 | 0.725 | 0.787 | 0.510 |

Table 4.7. Each model's performance on the *Age-Related* test set. The McNemar test indicated a marginally significant difference between each model's performance on the *Age-Related* test set, $\chi^2 = 2.78$, $p$<0.10. The test set contained 296 sentences sourced from blogs authored by older adults. The *Original* model classified 199 inputs correctly and the *Age-Removed* model classified 190 inputs correctly. The models disagreed on 23 inputs.

| Model | Non Age-Related Accuracy | F1 | Pos Precision | Neg Precision | Pos Recall | Neg Recall |
|---|---|---|---|---|---|---|
| *Original* | 79.61% | 0.812 | 0.763 | 0.841 | 0.870 | 0.722 |
| *Age-Removed* | 79.33% | 0.810 | 0.762 | 0.836 | 0.863 | 0.722 |

Table 4.8. Each model's performance on the *Non Age-Related* test set. The McNemar test indicated no significant difference between each model's performance on the test set, ($\chi^2 = 0.0$, $p$=1.0). The *Non Age-Related* test set contained 358 sentences sourced from Sentiment140.

as *positive* and the *Original* model rated the sentence as *negative*. Moreover, in the seven remaining disagreements, the *Age-Removed* model produced a *negative* outcome in reference to younger age. The *Age-Removed* model rated just one sentence referencing young age more positively than the *Original* model.

On the *Non Age-Related* test set the *Original* model very slightly outperformed the *Age-Removed* model, though the difference in accuracy was also statistically insignificant (79.61% vs. 79.33%, $\chi^2 = 0.0$, $p$=1.0). In fact, the model outputs diverged on just one test example in the set. The *Original* model produced *positive* and *negative* recall scores of 0.868 and 0.722, respectively, and an F1 metric of 0.812. The *Age-Removed* model produces *positive* and *negative* recall scores of 0.863 and 0.722, and an F1 metric of 0.809.

#### 4.1.6. Accuracy Evaluation Conclusion

My relatively simple manipulation of removing instances of "old" and "young" from the *Age-Removed* model training data set came with some cost to model accuracy as measured against older adults (aged 50+) as annotators. Although, the *Age-Removed* model's accuracy on the *Age-Related* test set was lower than that of the *Original* model, the results of the McNemar test indicate that differences in where the model's erred were only marginally significant. Differences in where the models erred on the *Non Age-Related* test set were also not found to be statistically different. The fact that the *Original* model performed with slightly better accuracy on both test sets suggests that the *Original* model may have better encoded biases– especially the age-related biases– of the test set annotators. This means that, depending on required accuracy, removing training data to create the *Age-Removed* model may not be a sufficient approach to dealing with age bias. In the next chapter I discuss advantages and disadvantages to this approach more in-depth.

Because the *Age-Related* test set consists of sentences and annotations derived from a source different from model training data, it is not surprising to observe diminished accuracy for both models compared to accuracy on the *Non Age-Related* test set. However, taken with measures of output bias, the *Age-Related* test set provides information about the extent to which a chosen model may be appropriate for measuring the attitudes of the population represented in

the test set– in this case older adults across the Unites States. The *Age-Related* test set is reusable for any sentiment models that produce "positive", "neutral", and "negative" outputs. Moreover, obtaining additional test annotations or annotations from a specific sub-population of older adults can be done at relatively little cost.

### 4.2. Study 2B: Older Adult Input in Training

My accuracy analysis introduces a first step in incorporating older adult input in model building. However, another approach to incorporating older adult input is to use older adults' annotations in model *training* rather than just testing. Soliciting input for model training builds upon my test set analysis in two ways.

First, a limitation of my *Age-Removed* approach of removing instances of "old" and "young" from the model's training data set is that social bias can take additional implicit forms. Stereotypical associations with "old" and "young" that are thematic may not be robustly removed by my approach, such as older adults and a lack of sexual interest or desire. Another limitation to my approach is that, in preventing the learning algorithm from associating discriminatory patterns with references to age, I also remove other, potentially unknown, associations with age. For logical reasons, older age is more strongly associated with 'retirement' and associated language than is younger age. While this may constitute a form of bias, it is not necessarily positive or negative, nor is it an inherent indicator of age-related social discrimination. These associations may be value-neutral or potentially important to the analyses at hand. Since training examples referencing age were removed altogether from the *Age-Removed* model, the model was hindered in detecting nuanced associations with aging that are actively discussed in the blog data set from which the *Age-Related* test set was sourced. In response to these limitations I ask whether soliciting annotations specifically from older adults might provide a way to preserve age-related associations while addressing age-related bias in sentiment models.

Second, while a test set of older adult annotations allows me to evaluate the extent to which a given model encodes older adults' views on age-related content, a training set provides an avenue of intentionally encoding older adults' views into a model directly. Whether or not these views are biased, they are necessary to encode if the goal of applying a model is to understand the true attitudes of the subjects under analysis. In addition, expanding the collection of older adult annotations to training data rather than just test data can help affirm if diminished accuracy from the *Age-Removed* model is the result of age bias exhibited by the *Age-Related* test set annotators or if diminished accuracy is an artifact of the *Age-Removed* model not having learned associations with age (even if biased). From the perspective of data representation, exploring older adult contributions to a training data set enables older adults to robustly shape how older age is represented and encoded.

### 4.2.1. The Annotation Process

Training data annotations serve as ground truth that indicate to learning algorithms how input data should be classified. Data and annotation quality is critical to the creation of well-performing models (Banko & Brill, 2001), which has led to the creation of gold standard data sets. Gold standard data sets, such as the Penn Treebank (Marcus et al., 1993), underpin much work across machine learning, especially natural language processing. They are commonly used to maintain comparisons in the training and evaluation of different models, and they feature large amounts of data typically annotated by expert sources.

At the same time, crowd sourcing has increasingly been used for annotation tasks because of its speed and low cost relative to the creation of gold standard data sets. With the rise of crowd sourcing and crowd work platforms, scholarship in machine learning has taken up comparisons of annotations provided by domain experts to those provided by non-experts (Snow et al., 2008). While crowd work platforms can provide valuable data, their worker populations are not very diverse and tend to hail from particular geographies (Posch et al., 2018), raising questions about the validity of crowdsourced data for applications involving underrepresented groups. Although researchers have raised questions about who comprises crowd worker pools, information about crowd work annotators goes largely unrecorded. As a result there is no standard practice for assessing who annotators are in relation to the data they are annotating. For example, I can deduce from general demographics of crowd work platforms that older adults are underrepresented, but it is rare that any data sets indicated the demographics of the workers who provided annotations. This is important to consider because individuals from different communities will interpret and annotate data differently (Patton et al., 2019; Sen et al., 2015) and even different gold standard data sets will produce different model results (Sen et al., 2015).

Consequently, it is important to be aware of who is represented in data annotator samples– especially for analyses involving subjective concepts. Important to consider in the present line of research is that older adults tend to be underrepresented on crowd work platforms, which means that any of their views and interpretations are near-absent from data sets created using crowd work. I purposely depart from typical crowd work annotation tasks and not only collect training data by soliciting older adults specifically, but also by collecting demographic and attitudinal data so I can better investigate how their perspectives shape annotation behavior.

I look to older adults' data annotations as a source for model training data (rather than test data) to create the *Older Adult* model. The architecture for the *Older Adult* model is identical to the previous custom models I created (i.e., maximum entropy, bag-of-words classifier). However, by collecting annotations directly from older adults, I attempt to

preserve known and potentially unknown associations with age while addressing output bias. Similar to my previous analysis of sentiment analysis tools, I assess the outputs of a sentiment model trained using older adults' annotations (i.e., *Older Adult* model) for both age-related bias and accuracy. I compare the *Older Adult* model's age-related bias and accuracy to the *Original* model and the *Age-Removed* model. Finally, I consider data annotations in relation to detailed demographic and attitudinal data collected from annotators so that I can precisely analyze how annotators' attitudes and beliefs influence their annotation behavior.

### 4.2.2. Method Overview

The first analysis I conduct assesses the *Older Adult* model for age-related bias— that is, its likelihood to classify sentences containing "old" more negatively than sentences containing references to "young" (*"I saw the old man at the movies"* vs. *"I saw the young man at the movies"*). This replicates the bias test I used in Study 1 (see **??**).

The second analysis tests the *Older Adult* model for accuracy on the *Age-Related* test set, duplicating the accuracy analysis just described. Once again, I defer to older adults as providers of the most relevant ground truth for age-related content in this context. As in the previous accuracy test, my goal is to assess the extent to which this approach to mitigating age-related bias might impact accuracy compared to using the *Original* model. I also repeat the accuracy test using the *Non Age-Related* test set, once again comparing results to the *Original* model. This analysis explores any impact on the the model's performance in non age-related contexts after replacing original annotations with those of older adults' to mitigate age-related-bias.

This set of analyses directly addresses whether a more inclusive model creation process might mitigate age-related bias while taking steps to improve data representation. I take a detailed approach to comparing a novel method of mitigating algorithmic bias to traditional methods of soliciting training data annotations and assessing model performance. The aim of this analysis is to provide an approach to ethically and robustly create a model and test its suitability in analyzing an underrepresented population. This work addresses whether data representation can be improved while maintaining model performance. This work also takes heed of who comprises the annotator population as well as their relevant demographics and attitudes.

### 4.2.3. Model Architecture & Data

The *Original* model in this analysis is the same *Original* model used in the previous analyses. The *Older Adult* model will differ only in the data used to train it. Using the same nationally representative panel through which I solicited test

| Model | Training Data Source | Training Data Size |
|---|---|---|
| Original | Sentiment140 training data | 1.6 million tweets |
| Older Adult | Sentiment140 training data + re-annotated subset | 1.6 million tweets |

Table 4.9. The sentiment models and training data sets.

set annotations, I solicit training data annotations from older adults (aged 50 or older) for 13,781 training examples from the Sentiment140 data set. In my previous analysis of custom sentiment models, I found that removing 13,781 training examples containing "old" and "young" was sufficient to mitigate observed age-related bias. I specifically obtain annotations on this data subset from older adults. By asking older adults to re-annotate these examples, my goal is to re-introduce these training examples so that the resulting model may learn age-related biases that align with those of older adults. Focusing on this subset of 13,781 tweets enables me to more efficiently leverage older adult perspectives on the most relevant data points while avoiding the large cost of re-annotating the entirety of the 1.6 million tweet data set. As I discuss in Chapter 6, one potential limitation to re-introducing age biases is that some uses may require equal treatment of age, such as for non-discrimination law compliance in employment contexts.

Although the available Sentiment140 training data set is annotated only for binary classification (i.e., 'positive' and 'negative' classifications), I solicit annotations from older adults on a 5-point Likert scale (Negative, Slightly negative, Neutral, Slightly positive, Positive). This is because the forced choice of *positive* or *negative* for examples that might otherwise be annotated as *neutral* may introduce noise and muddle the classifier's ability to discriminate between positive examples and negative examples. Similarly, including *neutral* as an annotation in model learning can improve sentiment classification compared to binary classification, as Koppel and Schler (2006) argue. In addition, I believe including a *neutral* category will be more intuitive to my respondent population, which may not be used to completing tasks typically administered through crowd working platforms.

In order to be directly comparable, both models must follow the same classification paradigm (i.e., both must produce the same sets of sentiment outputs). To generate a single ground truth annotation for each training data point, I again re-coded each ordinal annotation response on a scale from -2 to 2 and averaged the score. Each example with an average above 0 was recoded to "positive" and each example with an average below 0 was recoded to "negative". Any examples that averaged to exactly 0 were recoded to "neutral". To maintain the ability to compare the accuracy of the *Original* model to the *Older Adult* model, examples for which the annotator consensus is "*neutral*", of which there were 1,869, were excluded for model training. According to Koppel and Schler, removing "neutral" examples

helps to differentiate semantic features that contribute to *neutral* sentiment rather than other sentiment categories, removing noise and potentially improving model accuracy. Just as in the earlier accuracy analysis, examples from the *Age-Related* test set that are annotated as *neutral* were removed in order to maintain direct comparisons between the *Original* model and the *Older Adult* model.

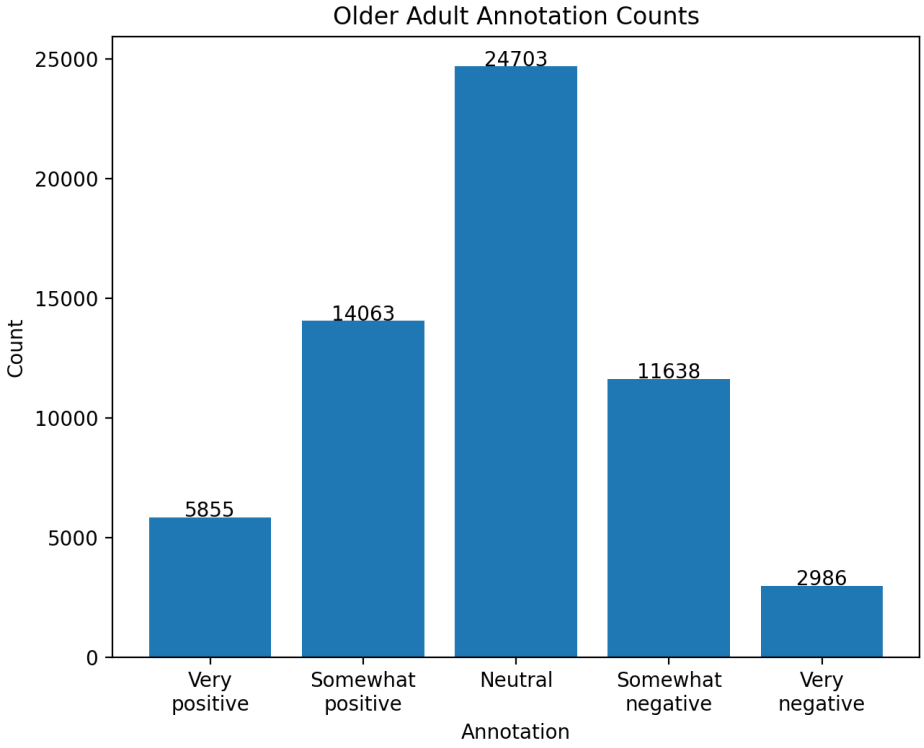


Figure 4.3. The distribution of sentiment annotations provided by older adults. In total there were 1,483 annotators.

### 4.2.4. Survey Data

To better understand the annotation population, the annotation task includes questions assessing respondent age anxiety and attitudes toward the use of age data in automated decision-making. Age anxiety describes the "combined concern and anticipation of losses centered around the aging process" (Lasher & Faulkender, 1993). In addition to standard demographic questions such as ethnicity and gender, the questions probing relevant attitudes and experiences with age are also included– for example, personal experience with or witnessing age discrimination, work status, and living situation. The survey, listed in the Appendix, also includes a 20-item survey assessing attitudes and anxieties about

aging borrowed from Lasher and Faulkender (1993). The survey features statements grouped under four constructs which together comprise the larger concept of Age Anxiety. The constructs are *Fear of Old People*, *Psychological Concerns*, *Physical Appearance*, and *Fear of Losses*, where each statement is followed a 5-point Likert agree/disagree scale.

Through these questions I gather a rich understanding of annotators' attitudes and experiences. The data allow me to compare attitudes toward aging to annotation patterns, as well as gauge ways in which older adults experience and perceive age differently. For this analysis, collecting data on annotation variability will add sensitivity to statistical tests. In this way, the data explicitly connects relevant attitudes about aging and lived experiences of annotators to observed algorithmic bias. Finally, for each data point annotators were asked to indicate whether the text prompt was "on the topic of age or aging". I do not use all of the data in the specific analyses I detail here, but I am publishing the data alongside this work for follow up research regarding connections between annotator attitudes, perspectives, and annotation behavior.

#### 4.2.5. Analyses for Age-Related Bias and Accuracy

In order to test the *Older Adult* model for age-related bias, I run the model on the *Age-Related* test set and first measure differences in model confidence for the sentences containing references to "old" compared with sentences containing references to "young". I conduct this test using a paired t-test where my independent variable is the relative age referent (i.e., "old" or "young") and the dependent variable is model confidence in a "positive" outcome.

To assess model accuracy, I calculate model percent accuracy, as well as model precision, recall, and F1 metric. I then run a McNemar chi-squared test to assess whether the *Older Adult* model outputs on each test set differ significantly from the *Original* and *Age-Removed* model outputs. I then repeat the McNemar test, this time comparing the *Older Adult* model to the *Age-Removed* model.

Finally, to gain a clearer understanding of the connection between the *Older Adult* model performance and underlying data, I conduct two more analyses. The first is a qualitative error analysis. I isolate the set of false positive and false negative sentences for the *Older Adult* model on the *Age-related* test set to thematically analyze the content of these sentences. Doing this highlights potential thematic similarities among sentences on which the model erred. I compare the content of these tweets to the set of false positive and false negative sentences for the *Original* model to understand similarities and differences in the types of sentences for which each model failed. I repeat this analysis on the set of false positive and false negative sentences on the *Non Age-Related* test set.

### 4.2.6. Possible Outcomes and Their Significance

One possible outcome of these analyses is that the *Older Adult* model will exhibit less age-related bias than the *Original* model. This is motivated by the fact that older adults have direct experience with age-related issues and may interpret age-related content in a way that does not draw heavily from ageist stereotypes. At the same time, it is possible that the *Older Adult* model will exhibit similar age-related bias compared with the *Original* model. This is motivated by research illustrating that social groups often internalize stereotypes about themselves (Levy, 2009). In this study, this means that older adult annotators may hold internalized, negative beliefs about older age and aging as a result of a broader societal preference for youth and younger age. Ultimately, I expect annotators to annotate based on their experiences and age-related values. Indeed, self focus bias described by Hecht and Gergle (2009), which stems from contributions that are important or "correct" to the contributing population, but not to other populations, may help to encode older adult perspectives in the trained model.

Based on the prior accuracy analyses, I expect that if the *Older Adult* model does exhibit age-related bias, its accuracy should be similar to that of the *Original* model. However, it is possible that the *Older Adult* model will produce *more* accurate outputs because older adult annotators differently rate training examples based on nuanced understandings of age that align with understandings expressed in the *Age-Related* test set. However, if the *Older Adult* model does exhibit not age-related bias, I expect its accuracy to be worse than the *Original* model, since the accuracy test.

### 4.2.7. Age-Related Bias Results

The results of the paired T-test indicate that the *Older Adult* Model differently treated sentences containing "old" from those containing "young". Sentences containing "old" were rated more negatively than those containing "young". On average, the model was 8.6% less confident that a sentence was positive when it contained the word "old" rather than "young" ($p < 0.0001$, $M_{old} = 0.580$, $M_{young} = 0.665$, SE = 0.008). This differences is near-identical to the *Original* model's performance in Study 1, in which the *Original* model was also approximately 6.7% less confident that a sentence was positive when it contained the word "old" rather than "young".

| Model | Age-Related Accuracy | F1 | Pos Precision | Neg Precision | Pos Recall | Neg Recall |
|---|---|---|---|---|---|---|
| *Original* | 67.23% | 0.70 | 0.624 | 0.746 | 0.787 | 0.568 |
| *Older Adult* | 65.88% | 0.707 | 0.598 | 0.793 | 0.865 | 0.471 |
| *Age-Removed* | 64.20% | 0.677 | 0.594 | 0.725 | 0.787 | 0.510 |

Table 4.10. Each model's performance on the *Age-Related* test set. The *Age-Removed* model is shown for comparison. The test set contained 296 sentences sourced from blogs authored by older adults. The Older Adult model classified 195 examples correctly and erred significantly differently from both the *Original* model ($\chi^2 = 23.31$, $p$<0.0001) and the *Age-Removed* model ($\chi^2 = 9.63$, $p$<0.002).

### 4.2.8. Accuracy Results

On the *Age-Related* test set, the *Older Adult* model performed worse than the *Original* model but better than the *Age-Removed* model. The importance of the relative differences in percent accuracy is dependent on the context of applications. Digging deeper into the models' patterns of classification, the McNemar test indicated a significant difference between the *Original* model and the *Older Adult* model with respect to the sentences each model classified correctly and incorrectly ($\chi^2 = 23.31$, $p$<0.0001). Applying the McNemar test to compare the *Older Adult* model to the *Age-Removed* model, there was also a significant difference in the error patterns ($\chi^2 = 9.63$, $p$<0.002).

Upon closer inspection, the *Older Adult* model had a tendency to rate sentences more positively than *both* the *Original* and *Age-Removed* models. This is reflected in the lower positive precision score (which indicates that, of all "positive" outcomes, relatively few were true positive), and the lower negative recall score (which indicates that the *Older Adult* model identified a relatively smaller proportion of true negatives). The *Older Adult* and *Original* models disagreed on 29 sentences, 28 of which the *Older Adult* model rated as "positive". Of those 28 sentences, 18 were in reference to older age and 10 were in reference to younger age. This means that the *Older Adult* model's tendency toward "positive" affected sentences referencing older age more so than sentences referencing younger age.

The *Older Adult* and *Age-Removed* models also disagreed on 29 test sentences, 23 of which the *Older Adult* model classified more positively. Of the 23, 15 referenced younger age and 8 referenced older age. All of the sentences that the *Older Adult* model classified more negatively were in reference to older age. To summarize, the *Older Adult* model had a tendency to classify sentences more positively than either of the other models; however, the *Older Adult* had a stronger tendency to classify old referents more positively in comparison to the *Original* model, but a stronger tendency to classify *young* referents more positively in comparison to the *Age-Removed* model.

| Model | Non Age-Related Accuracy | F1 | Pos Precision | Neg Precision | Pos Recall | Neg Recall |
|---|---|---|---|---|---|---|
| *Original* | 79.61% | 0.812 | 0.763 | 0.841 | 0.870 | 0.722 |
| *Older Adult* | 79.33% | 0.809 | 0.762 | 0.836 | 0.863 | 0.722 |
| *Age-Removed* | 79.33% | 0.809 | 0.762 | 0.836 | 0.863 | 0.722 |

Table 4.11. Each model's performance on the *Non Age-Related* test set. The McNemar test indicated no significant difference between the *Original* model's performance and the *Older Adult* model's performance ($\chi^2 = 0.0, p<1.0$). There was also no significant difference in error patterns between the *Older Adult* model and the *Age-Removed* model ($\chi^2 = 0.500, p=0.4795$). The test set contained 358 sentences sourced from Sentiment140. The models disagreed on just one input. The *Age-Removed* model is shown for comparison.

In addition, the *Older Adult* model produced an F1 metric of 0.707 as well as *positive* and *negative* precision scores of 0.598 and 0.793 respectively. The *Older Adult* model also produces a *positive* recall score of 0.865 and a *negative* recall score of 0.471. The relatively low *negative* recall score indicates that the model had a tendency to produce false negative outputs.

On the *Non Age-Related* test set, the *Older Adult* model performed slightly worse than the *Original* model. Although the *Older Adult* model and the *Age-Removed* model disagreed on two test examples, their overall percent accuracy was identical. Overall each model performed similarly on the *Non Age-Related* test set, with no statistical difference in their performance. The *Older Adult* model produced an F1 metric of 0.809 as well as *positive* and *negative* precision scores of 0.762 and 0.836 respectively. The model also produces a *positive* recall score of 0.863 and a *negative* recall score of 0.722.

### 4.3. Annotator Analyses

As a final analysis, I turn to studying annotation behavior of annotators I sampled to create the *Older Adult* model. I conduct both a qualitative and a quantitative analysis. The qualitative analysis entails isolating training and test set sentences that produce high disagreement among annotators, according to the metrics listed below. I then conduct a follow up qualitative analysis among these to investigate the origins of this disagreement.

In line with work that characterizes and measures annotation behavior in crowd worker tasks, I measure intersubjective agreement among my annotator population. Intersubjective agreement metrics provide insight into the general quality of the annotations solicited, as well as bring attention to the data examples and annotations on which annotators agree and disagree the most. Capturing disagreement is key for determining whether differences in annotation behavior are systematically related to annotator characteristics. To measure intersubjective agreement, I look to the CrowdTruth

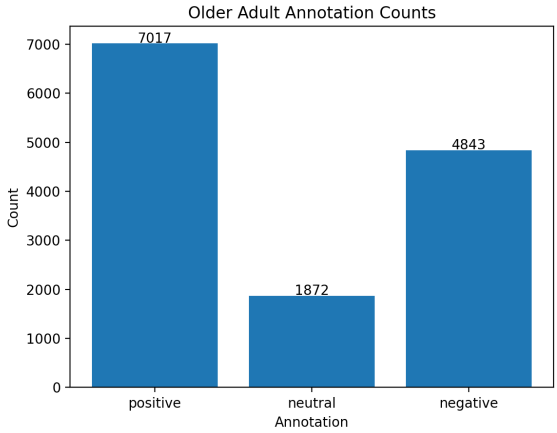


Figure 4.4. The distribution of "positive," "neutral," and "negative" annotations on the *Older Adult* training data subset featuring "old" and "young."

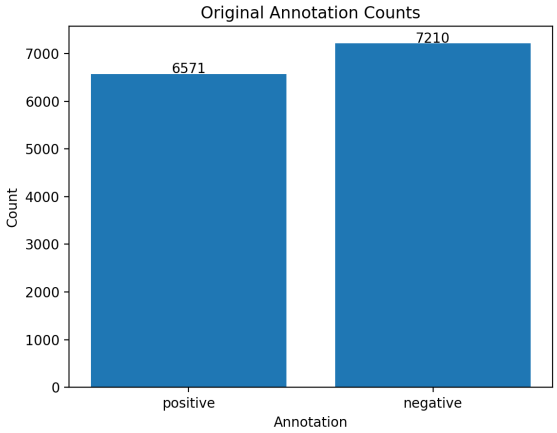


Figure 4.5. The distribution of "positive" and "negative" annotations on the *Original* training data subset featuring "old" and "young."

framework from Inel et al. (2014), which is specifically designed to understand annotator points of view and determine annotation quality in crowdsourced annotation tasks. Similar to my assertion that data sets created from human data will contain social biases, the authors reject the notion of a universal, gold standard data set. Through their framework, they characterize disagreement among annotators while acknowledging the mutual dependence of annotation quality, data point quality, and annotator quality. Of the metrics they describe, I use *Unit Quality Score* and *Annotation Quality Score*. Table 4.12 describes each metric.

As final tests, I focus on annotator survey responses in relation to annotation behavior. Driven by the demonstration by Sen et al. (2015) that different annotator populations will provide different ground truth labels on the same data, I

| *Metric* | *Description* | *Significance* | Score Range |
|---|---|---|---|
| **Unit Quality Score** | The average cosine similarity between all worker vectors, weighted by the worker quality scores and the annotation quality scores | Expresses overall worker agreement over one data point. | 0 to 1, where 1 indicates unanimous agreement among annotators |
| **Annotation Quality Score** | The probability that if one worker selects an annotation on a data point, another worker will also select it. | Measures the agreement over an annotation in all data points that it appears. | 0 to 1, where 1 indicates maximum agreement across all uses of an annotation. |

Table 4.12. The CrowdTruth agreement metrics.

seek to understand potential causes and correlates of annotators' ground truth determinations. I first look to annotators' age anxiety as a potential driver of age bias in annotation behavior. Then I look to demographic variables and annotator experience with age discrimination to generate a correlation table between survey responses.

### 4.3.1. Age Anxiety Tests

For each annotator I compare annotators' age anxiety to their annotations on age-related test sentences. The full distribution of annotator age anxiety is shown in 4.6. As a measure of respondents' concerns and stresses about aging and loss, I look to age anxiety as a predictor of how positively or negatively annotators view aging and age-related topics. I also tabulate demographic counts in relation to age anxiety to understand how age anxiety and views on aging may vary along with other demographic variables in my annotator sample.

I hypothesize that, as annotator age anxiety increases, the annotation on older age-related content will become more *negative*. Likewise, I use two t-tests to see if annotators with higher age anxiety have a tendency to annotate sentences referencing *younger* age more *positively*. To do this, I first split the *Age-Related* test set into one subset that references older age and another subset that references younger age. Because the test set is sparsely annotated, I cannot run a paired statistical test. Instead, I isolate all respondents who annotated the test set. In order to maximize

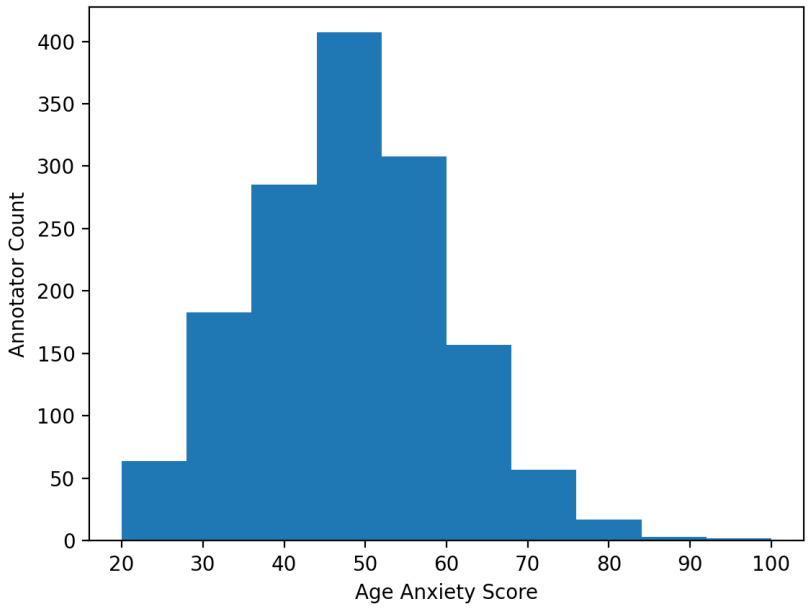


Figure 4.6. The distribution of annotator age anxiety.

| | High Anxiety Grp. | Low Anxiety Grp. |
|---|---|---|
| *Mean* | -0.16 | -0.06 |
| *SD* | 1.08 | 1.19 |
| *N* | 109 | 111 |

Table 4.13. The t-test results comparing anxiety group performance on the test sentences referencing *older* age indicate no significant difference between each group's tendency to annotate more positively or negatively ($p$<0.469, $M_{High}$-$M_{Low}$=-0.11, 95%CI= [-0.41, 0.19] SE=0.153).

the sensitivity of the statistical test, I group annotators into two groups based on age anxiety. I use z-score (or standard deviation distance) to group annotators for two reasons. First, because age anxiety in my annotator sample is normally distributed, there is no obvious age anxiety threshold to consider. Second, I want to preserve an *N* that is both relatively equal between the two groups, as well as large enough such that I do not compromise statistical power. I group annotators whose age anxiety z-score is greater than 1 into the *High Age Anxiety* group. Similarly, I group the respondents whose age anxiety z-score is less than -1 to form the *Low Age Anxiety* group. In the end, the *Low Age Anxiety* group consisted of 111 annotators and the *High Age Anxiety* group consisted of 109 annotators.

The first t-test compares the annotations of the *Low Age Anxiety* on the test set examples referencing older age to those of the *High Age Anxiety* group. The second t-test compares each group's annotations the test sentences referencing younger age. Important limitations to note in my approach are that, because test and training data were both sparse and randomly presented to respondents, some respondents did not annotate any test sentences and are, therefore, not represented in the t-test samples. Conversely, some respondents are represented more than once in within a sample. Finally, because of the randomized annotation prompts, *the anxiety groups in each t-test sample are not identical*. This means the *Low Age Anxiety* group in the first t-test does not represent the same group of annotators in the *Low Age Anxiety* group in the second t-test. This weakens the sensitivity of my analysis because, while I predict that age anxiety will influence annotation, I do not know the extent to which this effect might vary across individual annotators.

### 4.3.2. Annotator Analysis Results

Table 4.13 shows the t-test results for anxiety group annotations on test set sentences referencing older age and Table 4.14 shows the t-test results for anxiety group annotations. The results of both tests indicate that there is no significant difference between each anxiety group's tendency to rate sentences more positively or negatively than the other.

| | High Anxiety Grp. | Low Anxiety Grp. |
|---|---|---|
| *Mean* | 0.05 | -0.03 |
| *SD* | 0.93 | 1.11 |
| *N* | 96 | 129 |

Table 4.14. The t-test results comparing anxiety group performance on the test sentences referencing *younger* age indicate no significant difference between each group's tendency to annotate more positively or negatively ($p$<0.554, $M_{High}$-$M_{Low}$=0.08, 95%CI= [-0.19, 0.36] SE=0.140).

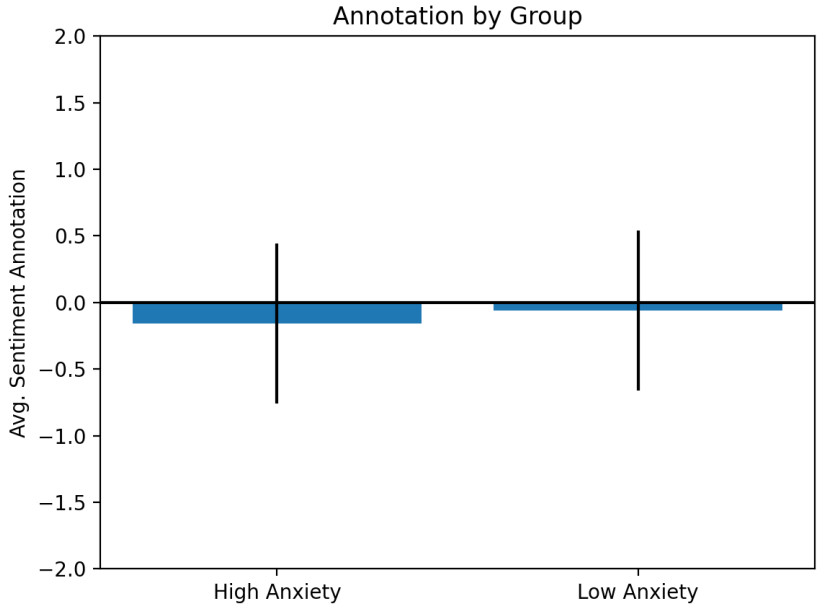


Figure 4.7. The bar graph shows the average annotation for each anxiety group (High= -0.16, Low =-0.06). The results of the t-test indicate that there is no significant difference between each anxiety group's tendency to rate sentences more positively or negatively than the other. Thus, age anxiety does not predict annotation behavior. The difference between each group was just .10, which is small relative to the Likert scale unit step of 1.

In addition to running the t-tests, I tabulated basic demographic counts for the High and Low Age Anxiety groups in each test. Although the t-test results are insignificant, data on how different older adults view aging differently serves to guide future analyses of age bias. Even if age anxiety does not predict annotation behavior, demographic differences related to anxiety may be connected to measurable biases related to aging. For the adults who annotated content referencing older age, the demographic breakdowns of each anxiety group are largely similar. However, two differences emerge. First, the low anxiety group features relatively more Black adults than the high anxiety group. In

addition, the low anxiety group also features relatively more adults living in the South, compared to the high anxiety group. This trend is identical for the adults that annotated content referencing younger age. This suggests that race and geographic region in the United States may be associated with differences in views on aging.

### 4.3.3. Agreement and Thematic Analysis

Finally, overall annotator agreement skews somewhat low, which may have an influence on the model accuracy I recorded. Low agreement may have roots in the population from which I sampled annotators, perhaps in the design of the annotation task, or even differences in how age is understood culturally between the present-day annotators and the authors of the training data and test data, which spans over 10 years. Although not certain, this may have influenced the lower accuracy of the *Older Adult* model on the *Age-Related* test set. Figure 4.8 shows a histogram of inter-annotator agreement by data point in the train and test data sets.

Figure 4.8 shows the distribution of annotator agreement across all training and testing set examples. Agreement skews low, raising questions about the content and quality of the data examples, as well as questions about the

| Race | | Gender | | Age | | Region | |
|---|---|---|---|---|---|---|---|
| *White* | 79 | *Female* | 60 | *50-59* | 58 | *West* | 32 |
| *Asian* | 18 | *Male* | 49 | *60-69* | 39 | *South* | 28 |
| *Black* | 7 | | | *70-79* | 9 | *Midwest* | 26 |
| *Native* | 4 | | | *80-89* | 3 | *Northeast* | 23 |
| *Other* | 1 | | | | | | |
| Hisp/Latino | 20 | | | | | | |

Table 4.15. The demographics of the High Age Anxiety Group that annotated the Age-Related test set sentences referencing older age.

| Race | | Gender | | Age | | Region | |
|---|---|---|---|---|---|---|---|
| *White* | 80 | *Female* | 58 | *50-59* | 35 | *West* | 20 |
| *Asian* | 7 | *Male* | 53 | *60-69* | 46 | *South* | 57 |
| *Black* | 20 | | | *70-79* | 29 | *Midwest* | 17 |
| *Native* | 2 | | | *80-89* | 1 | *Northeast* | 17 |
| *Other* | 1 | | | | | | |
| Hisp/Latino | 19 | | | | | | |

Table 4.16. The demographics of the Low Age Anxiety Group that annotated the Age-Related test set sentences referencing older age.

potential challenges of working with non-traditional annotator populations. Compared to high annotator agreement, low agreement means that a learning algorithm will be presented with an unclear signal from which to learn patterns and associations with classification categories (i.e., "positive" and "negative"). In the absence of a clear signal, a trained model is more likely to produce error. Therefore, low agreement is one potential factor contributing to lower accuracy on the *Age-Related* test set.

| Annotation | Quality Score |
|---|---|
| Very positive | 0.310 |
| Somewhat positive | 0.455 |
| Neutral | 0.691 |
| Somewhat negative | 0.456 |
| Very negative | 0.249 |

Table 4.17. The annotation categories and their quality scores.

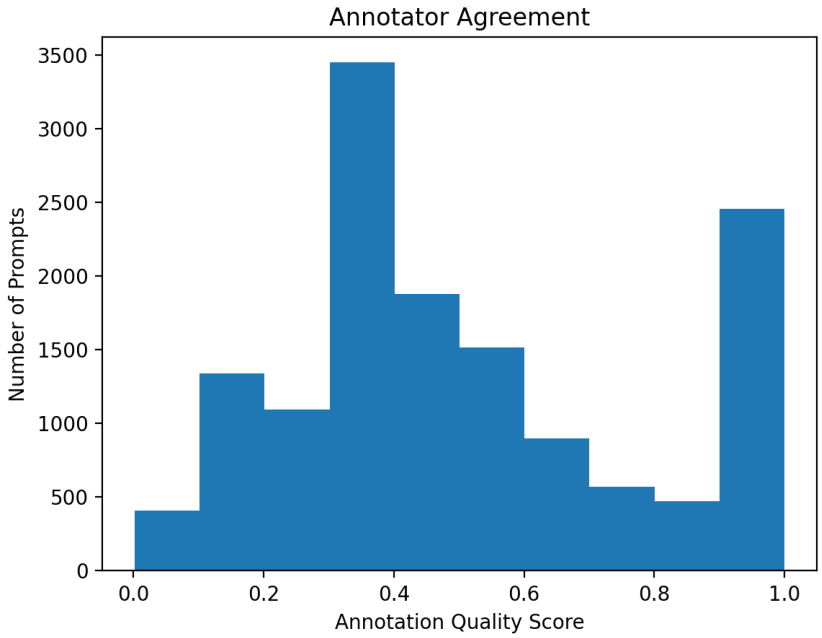


Figure 4.8. The distribution of annotator agreement across all annotated examples.

### 4.3.4. Qualitative Analysis of Agreement

In my follow up thematic analysis I focus on the 300 training and testing examples with the highest agreement and the 300 examples with the lowest agreement according to the CrowdTruth agreement metrics. In the analysis, my goal is to surface potential causes underlying annotator agreement and disagreement to guide follow up work. Before conducting the analysis, I collapsed the annotation categories. I combined "Very positive" and "Somewhat positive" into "Positive, and "Very negative" and "Somewhat negative" into "Negative". I collapsed the annotation categories because I was primarily interested in agreement regarding sentiment rather than the agreement in the relative magnitude of sentiment.

Unsurprisingly, many of the data points with highest agreement frequently featured unambiguous signals of sentiment or emphasis, including words or phrases of emotion such as "love", "hate", or "I'm so bummed", often alongside exclamations (e.g., *"@ArteDeb that's a big old stump! Love it!! How are ya' Debbbb?"*, *"watching old moveis is the best James Stewart, Katherine Hepburn & Cary Grant!"*, *"Feels like crying. I'll be done with my kiddos in 2 1/2 days, my sister is graduating, andni feel old and fat"*). These indicators of affect seemed to drive "positive" and "negative annotations. "Neutral" annotations often occurred in two contexts– when there was no obvious affect indicated in the text prompt, as expected, and when there was more than one sentiment or a complicated sentiment indicated. Text prompts with no obvious affect were those that reported information or had little context (e.g., *"Quick stop in Hopland @ the Mendocino Brewing Co - Cali's oldest brewpub apparently...no dogs allowed"*, *"trying to save my old post"*, *"Today is pocket money day, as they are still very young have said 5 each a month, performance related .. so this month only 3.50 each"*.

Text prompts containing more than one clear sentiment sometimes had conflicting sentiments (e.g., *"People I didn't even know were dating people from HS/College are now engaged/married. Is it possible to feel very old and young at once?"*, *"@minnymichelle omgod that dog! lol I miss my little red plastic gun - sighhh I think old school nintendos are expensive now."*. In this instance it is unclear if "neutral" was selected by any individual annotator because it represents a possible "average" sentiment across the whole tweet, or if "neutral" represented a catch-all for anything that was not apparently "positive" or "negative". In the particular annotation schema used in Sentiment140 and my custom models, multiple or complicated sentiments could introduce error; however, may not be an issue for schema that include a broader array of annotations or allow multiple annotations to be selected for a single data point.

At the same time, there was variation in whether annotators seemed to annotate in response to the likely valence of information being reported rather than the emotional state of the author. For example, the prompt "@herathcrush: its a

girl. They have 4 girls now The 2 oldest are teens and the 3rd is 7 yrs old." was annotated by all 4 annotators as positive. The tweet describes factual information about children's ages; however, the information seems to be in response to childbirth, which is widely considered to be a happy event. Similarly, "playing with my 14 week old daughter" was annotated as positive by all 4 annotators, despite a lack of information about the author's opinion on the activity.

Among the training and test examples with lowest annotator agreement, many examples seemed to have clear valence, whether in relation to the author's emotion or the valence of the content (e.g., *"on the news somebody killed a 5 yr. old girl! how could they!"*, *"How could anybody put a gun to a 9 yr old girls head & pull the trigger? The news completely sickens me anymore"*, *Cleaning old house for the last time. Joy*). It was unclear if annotations were driven by specific content described in the tweets or if annotations were affected by factors unrelated to the specific tweets, such as task fatigue or accidental annotation selection. Although I collected a variety of information about annotators in addition to their specific annotations, I did not collect information related to potential task confusion, fatigue, or difficulty with specific text prompts. Given the targeted nature of both the annotator population and subset of data for annotation I selected, follow up work could more deeply probe reasoning and motivations behind annotation decisions. Think-aloud protocols, surveys, or even interviews focused on a small number of annotation prompts could elucidate how annotators interpret task prompts. Relatively little work focuses squarely on open-ended approaches to surfacing agreement issues rooted in potentially systematic annotator confusion or social views, however Aroyo and Welty (2015) outline common problematic assumptions about the validity of human annotations, and Mohammad (2017) highlights challenges specific to language and sentiment annotation task design. In addition to age anxiety, which I have explored in this study, views shaped by social movements or marginalized experience may surface in annotation behavior in non-obvious ways.

Notably, the instructions for my annotation task did not offer annotators any specific definition of sentiment to use (i.e. "Please indicate how positive or negative the following text is."). I did this for two reasons. First, there is no standard convention for defining to annotators how sentiment should be interpreted, and, second, providing a definition can bias annotation behavior (Mohammad, 2017), which could have potentially obscured ways in which older adults exhibit similar or different degrees of age bias in their annotation behavior. Mohammad (2017) discusses challenges in sentiment analysis regarding identifying or annotating the valence of the speaker's emotional state compared with the valence of neutrally-reported information (e.g. "my mom died"). Further research, however, can leverage this insight to test attempts to improve agreement through unambiguous annotation task instructions.

CHAPTER 5

# Responding to Social Bias

Studies 1 and 2 chronicle a series of analyses that characterize age bias in sentiment analysis models and introduce approaches to identifying and mitigating it. A major step toward addressing age bias is acknowledging that computational technologies are not impervious to social biases in society. An obvious, recurring motif throughout my work, this acknowledgement cannot be understated. Acknowledging the links between social bias in computational models and society reframes social bias from something universally bad to be removed or deleted from data to something with roots and consequences we can robustly investigate and shape in service of stakeholders. As a result, I call on researchers and engineers to investigate biases that emerge to better understand their implications in application.

## 5.1. Implications of Age Bias

The existence of age bias in language modeling carries particular implications for older adults and underrepresented populations more generally. I purposely constructed my analyses to emulate realistic research contexts. Not only did the contexts emulate the initial research in which I attempted to use computational tools to study older adult bloggers, but they also recreated common research contexts in which non machine learning experts might adopt readily available algorithmic tools to conduct their work. The sentences I used to test for age bias and for the *Age-Related* test set were produced by a community of older adult bloggers discussing aging experiences and come from a real site of academic inquiry. Discussions cover a wide range of topics, such as politics, health, government, pop culture, and news, in relation to older adult experience. Thus, when the sentiment analysis models are applied to understanding the views, opinions, and experiences reported in this corpus, sentiment output is found to be less positive simply because the sentences describe an older person taking part in the interaction. For example, the statement "This old guy was 3 or 4 feet from the tide line and the tide was going out" was rated less positively than the sentence "This young guy was 3 or 4 feet from the tide line and the tide was going out." This is problematic when examining sentences from this corpus that may be mined by algorithms to understand attitudes towards products ("I love seeing older non-professional women modelling clothes."), health information ("The older adults' brain scans showed activity in the same area"), and learning ("Life and learning does not end in old age").

Beyond my specific research context, decisions by researchers and companies can be influenced by the relative sentiment of older adults' experiences compared to younger people, potentially affecting the products and services that are invested in and that are made available to older adults. Domains impacting older adults include analyses of public response to Medicare and older adult healthcare and reception of products and services targeted toward older adults. It is particularly noteworthy that I was able to detect age bias in my chosen sample sentences, which feature opinions and articulations of aging experience that are specifically crafted to challenge negative mainstream ideologies surrounding older age. My prior qualitative work analyzing samples the broader blog corpus emphasized this aspect of communication and the definitional role that online communication played in community development of an anti-ageist agenda. Blog posts explicitly described older adulthood as complex and fulfilling in ways not reflected by popular stereotypes (Lazar et al., 2017). However, researchers using sentiment analysis to understand expressed attitudes across the lifespan would find that statements describing older adulthood (e.g., "Every one is telling the world what it's really like to get older") are inherently less positive than those describing youth, even when this language is designed to promote positive associations with growing older. This sheds light on the extent to which social bias present in training data can overpower signals in test or research data.

### 5.1.1. Challenges of Studying Social Movements and Underrepresented Groups

In addition to general challenges related to social bias present in data, my analyses speak to challenges linked to the study of underrepresented groups and minority opinion. Using computational techniques to study social movements and the emergence of non-dominant narratives situated within particular cultures is becoming increasingly common (e.g., (Starbird & Palen, 2012; Twyman et al., 2017)). However, a central aspect of social movement formation involves using language strategically to destabilize dominant narratives in society and calling attention to underrepresented social perspectives. That is, language use is changing and evolving along with the emergent social movement. The shifting terms and specialized uses of language present a clear challenge for sentiment analysis models trained on static data that do not reflect evolving attitudes and language usage.

While the data I used to create my custom sentiment models was collected in 2016 and supplemented even more recently through my panel survey, aspects of the data set will likely lose validity with time as *both* prevailing and minority views on aging shift and change. This adds a layer to the already difficult task of capturing underrepresented experience in data. For example, my qualitative analysis of older adult bloggers showed that, as part of forming a social movement recognizing ageism, they reframed the aging experience as a positive and natural aspect of life. One

way in which they did this was by reclaiming words related to age that may have negative connotations, such as "gray," to have positive or alternate meanings (e.g., "Go gray!" is a common phrase used within this blogging community to positively promote allowing natural changes in hair color with age). The misalignment between static training data sets and evolving contemporary language makes understanding bias increasingly complex. Important to note is that the anti-ageist views shared among the community of bloggers is not representative of all older adults, such as those in my general annotator sample. This means that the language they re-appropriate and invent may not be used significantly among other adults, let alone people of various ages in broader society. At that, detecting and disambiguating novel use of common terms, slang, and language appropriation is a distinct area of research complicated by the fact that these terms cannot easily be anticipated or annotated at the their initial emergence (ElSahar & El-Beltagy, 2014; Matsumoto et al., 2014).

While there are challenges to applying computational techniques to text-based corpora, one advantage is the ability to observe phenomena of bias at a societal level that may be difficult to detect on an individual or case study basis. In this way, computational models can stand as a barometer to track shifts in social bias over time and across different contexts. Within research on aging, tools such as the Implicit Association Test (Greenwald et al., 1998) and Measurement of Aging Anxiety (Lasher & Faulkender, 1993) are used to assess attitudes towards older adulthood (e.g., (Hummert et al., 2002)). While these tests are useful for understanding individual attitudes, they probe overtly about issues of age bias and would require an extremely large sample to understand broader views on aging in any specific cultural context. In contrast, sentiment analysis models can be a lens for understanding underlying bias that is difficult for humans to detect overtly themselves. Indeed, age-related bias is a global phenomenon that has until recently been largely neglected in discussions of social justice and equality (Butler, 1969; Officer et al., 2016). Sentiment analysis and other computational tools can be valuable barometers to understand broader attitudes towards various social dimensions of society, such as aging. The supervised learning approaches I used had the capacity to measure and reproduce age bias from my annotator sample, as evidenced by the similar levels of bias among the *Original* and *Older Adult* models. Perhaps more importantly, knowing the social biases encoded in test data sets and their underlying annotators means that accuracy measurements from test sets can be used as litmus tests for potential alignments in social bias. That is, the higher *Age-Related* accuracy of my *Older Adult* model compared with my *Age-Removed* model suggests alignment between age bias encoded in the *Older Adult* model and *Age-Related* test set. Computational models and data sets have

already been shown to encode a slew of social biases that researchers likely have interest in tracking across different domains (Bolukbasi, Chang, Zou, et al., 2016).

## 5.2. Removing or Preserving Bias

Throughout my studies, I have explored various methods of identifying, measuring, and responding to social bias in model data. Given the range of approaches that I and others have explored, a natural question is how to determine the best approach to mitigate unintended bias in a given context. Moreover, what should happen when technical limitations prevent that goal from being met? At the same time, a critical decision in the design of algorithmic systems concerns when *not* to mitigate bias and instead preserve human biases as they truly exist. Mainstream discussion and news headlines on social bias in algorithms have largely framed the presence of social bias as wholly negative. Indeed, social bias in algorithmic systems can have devastating impacts. However removing it may not be always be an appropriate option and there may be limits to doing so.

### 5.2.1. Bias Removal

My straightforward method of removing age-related examples from a training data set stands as one way to force a resulting model to treat young and older age similarly. While accuracy should be evaluated in specific application contexts, my analyses suggest that the cost to overall accuracy is not significant. Removing age-related examples makes sense for contexts in which we want to counteract known, problematic human biases that should not influence a decision-making process. For example, in analyses such as automated resume screening, it may be prudent to specifically ignore or remove human social biases that stand to discriminate against particular social groups. Aside from normative beliefs that discrimination based on certain social identities is immoral, compliance with labor laws places limits on the extent to which social identities can play a role in assessment. In this context, removing socially biased data forces a chosen behavior that is socially and legally compliant, even if that behavior is not generally reflective of mainstream attitudes and beliefs. However it is worth noting that legal compliance in and of itself is not a definitive indicator of social bias since employment-related age bias is still prevalent and social discrimination can be challenging to adjudicate (Crenshaw, 1990). My bias removal approach may still reduce the degree to which algorithms can encode social bias such that social bias in outputs is also reduced or even removed altogether. In my manipulation, removing a relatively small number of data points also removed any valenced signal with respect to “old” and “young”. Even though “old” and “young” are not all encompassing descriptors of age, their removal from the training data set

significantly reduced bias in model performance. It is worth noting that I tested for just one form of fairness, equality of odds– or an equal distribution of outcomes for all tested input categories. If other forms of fairness are desired, the approach I took may not be appropriate for removing bias. For example, it may be desirable for outcomes across groups to match their population proportions in a data set, such as when considering racial equity in health outcomes. It is also important to acknowledge that sensitivity to underlying social bias in data may vary between the maximum entropy classifiers I tested and other models such as those based on neural networks.

Though seemingly a simple manipulation, removing data is only straightforward if bias can be attributed to a discrete input, such as the specific words "old" and "young". As I previously noted, isolating implicit or thematic associations, such as women with domestic work, is more complicated. Indeed, my analysis of word embeddings also demonstrated evidence of age bias in words implicitly associated with age. This suggests that simply removing instances of particular word tokens would not wholly remove all instances of a particular social bias in data. Work by Gonen and Goldberg (2019) similarly outlines vestiges of systematic social bias left in word embeddings after employing de-biasing methods. Removing word tokens also becomes increasingly challenging or impossible when considering the interplay of ageism and other forms of discrimination such as sexism and racism.

Another important consideration while using a data removal method to reduce bias is that it forces fairness of equality in outputs only with respect to model features and inputs that we can isolate and quantify. Though seemingly obvious, this underscores a need to explicitly report fairness and accuracy outcome of systems. Reporting for my *Age-Removed* model, then, should indicate that the model performs with reduced bias *with respect to sentiment treatment of the words "old" and "young"*, rather than reporting that might suggest the model is wholly free of age bias. Indeed, critical race theory explicates how social bias permeates through nearly all aspects of life (Delgado & Stefancic, 2017) and researchers in HCI have called out a need to acknowledge how social biases such as racism similarly permeate the digital world (Dietrich, 2013; Ogbonnaya-Ogburu et al., 2020).

### 5.2.2. Re-Biasing

An alternative to removing measurable evidence of social bias is to intentionally imbue a model with social biases that align with those of end users or other stakeholders. To ensure that a given model accurately measures the behaviors and attitudes of a target population requires the values and biases of the target population be encoded in the model. That is, to accurately characterize audience responses to a new product or service, the values and biases of that audience must be accounted for. However, it is the specific social biases of a *particular* audience or stakeholder that must be encoded

rather than an 'average' of all social biases or specific social biases expressed by other groups. This is not to say that discriminatory values should be purposely incorporated into decision-making processes or that final decisions should necessarily be left to automated processes. The aim is to encode the value system of a defined target population whose attitudes are the subject of measure. Study 2 took aim at "re-biasing" a sentiment model for three reasons. First, with a focus on age-related content, older adults were the most relevant audience to test against. Second, re-annotating data was an initial foray into mitigating unwanted bias while preserving other age-related associations in text. Finally, by recording input from a specific annotator population while additionally collecting survey and demographic information I was able to better trace potential roots of social bias.

In the end, however, adding older adult annotations to the training data re-introduced age bias that I had previously reduced. My analyses suggest that the older adults in my sample re-produced internalized age bias. This complicates my initial prediction that soliciting input from older adults could reduce unintended age bias. In order to counter act social biases embedded in data sets, it may be necessary to more specifically sample annotators whose social views explicitly stand in contrast to social biases at play. Internalized biases that annotators may possess also call into question whether additional effort to access a non crowd worker population is worth expending to produce a model that still exhibits age bias. Allowing stakeholders to shape system values remains a critical component of design rooted in values. My results suggest that my approach of collecting training annotations is not sufficient on its own if the goal is to reduce age bias. Asking annotators to provide novel training examples in addition to annotations may help a model more strongly encode the values of a chosen population.

Importantly, the *Older Adult* model McNemar test on the *Age-Related* test set indicated that the *Older Adult* and *Original* models had a significant tendency to err differently, despite the fact that both classified references to older age more negatively. Moreover, this test result was much more significant than the McNemar test comparing the *Original* model to the *Age-Removed* model, despite the fact that the *Original* model and *Older Adult* model performed with closer percent accuracy. This suggests that the age bias exhibited by the *Older Adult* model was qualitatively different than that of the *Original* model. Indeed the *Older Adult* model had a tendency to classify sentences more positively compared with the *Original* model– particularly those referencing older age. Collecting additional insights from older adults regarding annotator disagreements or asking older adults and follow up thematic analysis of false positive and false negative classifications, might provide more nuanced insight to the nature of each model's age bias.

Ultimately, both the *Original* and *Older Adult* models exhibited bias. Importantly, I drew annotators from a general U.S. sample of older adults according to race, gender, and geography, and my bias results may have differed had I sampled older adults with social and political views mirroring those expressed in the blog discussions from which I scraped my *Age-Related* test set. Although it has been several decades since ageism was first defined and popularized (Butler, 1969), survey work by Braithwaite et al. (2002) demonstrates somewhat counter-intuitively that high awareness of ageism in society among a sample of older adults was poorly correlated with a tendency to reproduce ageist beliefs. In other words, being aware of ageism did not make respondents less likely to reproduce ageism themselves. From a values standpoint, this means that any effort to intentionally design a sentiment model to reflect anti-ageist values would likely require explicit sampling of annotators who espouse these views. In fact, such explicit sampling need not be narrowly restricted to older adults. At the same time, however, imbuing a data set with nuanced associations with aging and older adulthood points to a need to preserve a significant sample of annotators with first-hand experience with older age.

The unintuitive relationship between ageism awareness and attitude points to the importance of social and political viewpoints in shaping opinion. Ultimately, it is important to acknowledge that there is not some inherent quality to older adulthood that shapes annotation behavior in one specific way. Although my sample consisted exclusively of older adults, their social and political views, like those among members of any social identity group, are not monolithic. While certain points of view may be more prevalent among older adults, the influence of ageism in society means that a significant number of older adults espouse negative views on aging that reflect discriminatory views upheld by society. As a result, efforts to sample from marginalized stakeholders necessitate decision-making about whose values to prioritize, among the broader marginalized group. Had I only sampled annotators with anti-ageist views, I expect that model bias would have been different from that of my *Older Adult* model. I would not expect measurable age bias to disappear, as with the *Age-Removed* model; however, I would expect age bias to shift in a way that treats references to older age less negatively relative to younger age. I would also expect shifts in age bias with respect to content describing older age in relation to topics typically interpreted more negatively, such as older age and sex or older age and beauty.

This makes decisions about selecting annotators complex, even within the narrow scope of age bias in sentiment analysis. Although I tested the influence of age anxiety on annotation behavior, it did not prove to be significant. However, framing annotations as an artifact of interpretive lens leaves open opportunities to study other aspects of

social views and the ways in which they may shape annotation behavior. It is precisely for this reason that my survey to annotators includes questions related to first-hand experiences with age discrimination and whether respondents identify as older adults. In addition to age anxiety, factors worth considering in relation to annotation behavior include whether or not annotators identify as older adults, the extent of their first-hand experience with age discrimination, and the extent to which they believe age discrimination is a problem in society. Considering other sociodemographic factors that can influence the nature of age bias, gender, socioeconomic status, and race/ethnicity may also prove fruitful to investigate with respect to how specific content themes are treated by sentiment models. Although I did not specifically investigate these factors in the present dissertation, I collected this data for ongoing and follow up analyses. Investigating bias with respect to these social identities may reveal unique ways in which they connect to age bias. At that, understanding anti-discriminatory ideologies invested in the liberation of these social identity groups (e.g., antiracism, feminism, economic justice) can reveal who should be consulted and involved in data sourcing, annotator sampling, and strategic evaluation toward equitable systems.

Annotators acutely aware of ageism in society may annotate in response to societal context in unique ways. For example, an anti-ageist annotator might rate the sentence, "We live in a culture that deliberately hides and ignores older folks." more negatively than other annotators due to personal disapproval of the action, "ignoring older folks," rather than because they interpret the tone of the statement to be negative. This underscores that negative treatment of older age, alone, is not a singular measure of discriminatory age bias. In addition, age is not an isolated social phenomenon. Rather, individuals' experiences with age influence and are influenced by other social identities. Prior research assessing age bias highlighted women's discussions of the ways in which intersections of sexism and ageism in society produce unique challenges for older women (Lazar et al., 2017). Similarly, C. N. Harrington et al. (2019) carefully chronicle health needs that must inform tailored solutions to address health disparities experienced by older adults who are Black and low-income. Beyond internalized beliefs about aging, folks can internalize discriminatory beliefs about other stigmatized identities they hold. Therefore, I expect internalized negative attitudes about a wide range of social identities to play a role in data set creation for computational models as well. For this reason, it is important to consider not just how a target stakeholder differs from other groups, but also diversity within the stakeholder group.

Interestingly, I found demographic differences between the High and Low age anxiety groups, wherein annotators with low age anxiety had a higher tendency to be Black and living in the southern United States. Although I did

not robustly test these demographic differences, they raise questions about whether Black individuals and individuals living in the South hold more positive views on aging. The research motivating my decision to measure age anxiety does not deeply discuss race or geographic region in relation to views on aging; however, future work should consider if these factors relate to age bias. The survey and demographic questions I deployed alongside my annotation task serve as an open-ended probe of differences in how older adults view and experience aging.

Another issue influencing bias is the context in which and for which training data was originally generated. One possibility for the bias observed in both the *Original* and *Older Adult* models is that the approximately 14,000 training examples containing "old" and "young" were imbalanced in how "old" and "young" appear. That is, the word "old" may have more often appeared in sentences widely agreed to be negatively-valenced, such as insults. Even if older adults possess views on aging that differ from mainstream stereotypes, they may likely agree with other annotator populations that the examples are indeed negative. Such age bias embedded in training data has roots in whose voice is represented in the source of training data, rather than its annotations–or interpretations. Ensuring a range of age-related training examples, then, is necessary for drawing out nuanced age representation. However, social systems impact who feels welcome to engage on a digital platform and which activities they feel comfortable engaging in, discussions of data source must consider the values encoded in the structures that shape data production. Indeed the designs of popular social networks may cater to users who are younger, leading older adults to engage with different platforms despite an interest in popular sites (Xie et al., 2012). Users who are implicitly discouraged from engaging with a system will not be robustly represented in platform data. This means that, despite older adults' active content production online (Waycott et al., 2013), rich characterizations of their experiences may be altogether absent from data sets if algorithm designers are not careful with data selection. Engineers and researchers creating algorithmic systems must identify and consider the relevant social views of individuals involved in creating source data and data annotations.

Collecting and reporting information about annotator and data source can aid in tracing social biases. My approach builds on Nissenbaum's (2001) call for researchers to conduct detailed, technical investigations of computational systems. That is, in addition to technical investigations of how models operate, researchers and engineers must also investigate and articulate the characteristics and producers of the data they use. The differences in age bias among the models I created suggest that, beyond the decision to preserve or remove age bias, there is also a question of *which* age bias to preserve or remove for a given application context. This speaks to auditing social expressions and traces embedded in data sets that Gillespie (2014) describes. Identifying these expressions and traces may help illuminate

potential biases in model performance. It may also reveal what is gained or lost by preserving certain biases in a model over others.

Detailing and reporting this information also raises questions about user and stakeholder trust based on this new data. As demand increases for algorithmic transparency, researchers must also consider how end user trust might be impacted by sources of training and validation data. Regardless of who is selected to provide data and annotations, the source is important to catalog in relation to end-user and stakeholder trust. Recent research in HCI has shown that algorithmic interface transparency can have negative impacts on trust if there is too little *or* too much information, particularly when algorithmic systems violate end user expectations (Kizilcec, 2016). A system built on data from a population whom an end user trusts may impact their trust in the broader system compared with a system built on data from a population whom an end user does not trust.

### 5.2.3. Cost & E□ciency

In addition to considering techniques to identify social bias as well as de-bias or re-bias models, the cost of implementing them and the availability of data will invariably influence their feasibility. Although field studies like the interviews I conducted to assess walkability provide rich and valuable information, they come at the cost of additional time and money. This cost stands to increase for studies involving difficult-to-access populations. At the same time, open-ended explorations may be incorporated into existing need-finding practices or adapted through surveys or other, more structured methods. The advantage of field methods is not a comprehensive audit of potential biases that might emerge when using a system within a given community, but rather rich understandings of values and biases in addition to those that are readily measurable quantitatively. In this way qualitative methods complement quantitative practice so that algorithm designers can identify system limits and debug their operationalizations of target concepts. For subjective concepts, qualitative methods may ultimately save time and effort because they can be undertaken before prototyping or model building. As a result, the costs of collecting new data or re-annotating data to reduce model error may be avoided. Still, collecting additional data annotations may not be a prohibitive cost. The cost to collect additional annotations for both the *Older Adult* model and the *Age-Related* test set was $4,236. While likely more expensive than collecting annotations from a crowd work website, using a survey panel service afforded me greater ease of access to a representative sample of my specific population of interest. At that, the cost is relatively low for many companies and research organizations. Researchers and engineers investigating fairness with respect to specific stakeholder groups should prioritize data from those groups– whether in the form of test inputs or data annotations.

By design, methods I explore in my sentiment analyses save time and cost compared to building new data sets from the ground up. In the creation of a brand new data set, my bias removal approach might initially seem inefficient due to tossing out data that might provide other kinds of potentially valuable information to a learning algorithm. However, this approach becomes much more useful when re-purposing the data set for additional applications or adapting a publicly available data set. Both the *Older Adult* and *Age-Removed* models adapt an existing data set by removing or re-annotating less than 1% of the total data. In this way, existing data sets can be better adapted to specific research contexts at relatively little cost, expanding the number of viable computational tools for research focused on populations with limited or hard-to-access data. In my analyses, the only cost I incurred was the relatively small amount of time it took to search through and remove training examples before building a new model.

### 5.3. Testing

Finally, test data is a valuable site for probing bias. In tandem with generating training data, testing data is important for not only identifying bias but also for evaluating proprietary algorithms. Proprietary algorithms prove to be a concerning issue in regard to many new technologies. With technology policy lagging behind technology development, many companies are left to self-govern around issues of social bias. The result is limited checks on fairness that can produce significant social bias in deployment (Buolamwini & Gebru, 2018; Larson et al., 2016). Proprietary technologies force researchers to take approaches similar to my initial exploration of age bias in sentiment analysis. Because details describing algorithm design are often protected by intellectual property law, researchers must analyze inputs and outputs with limited information about algorithm and data processing or take best guesses at recreating existing models. Policy aside, making use of publicly available tools responsibly can be aided through the creation of test data sets. Observing how a model behaves on given test inputs becomes a primary means of evaluation.

My *Age-Related* test set is an example of how we might develop test sets to address challenges presented by proprietary algorithms. Creating a test data set that is representative of a particular stakeholder group and information-rich with respect to stakeholder attitudes and values enables targeted evaluations. I used older adults' annotations of test data as a litmus test for sentiment models' alignment with older adults' interpretations of age-related content. My age-related accuracy analyses help to provide information about the fitness of a provided sentiment analysis model for analyzing older adults. The same can be done for any number of relevant stakeholders. One caveat to this approach is that, relevant stakeholders may still produce socially biased data that perpetuates discriminatory beliefs about themselves. The analyses of my *Older Adult* model suggested just this. Accurate performance is not the same as

unbiased performance; however, high accuracy might suggest an alignment of values. Using this approach, researchers focused on fairness in machine learning can use targeted annotation approaches or subset their annotation data sets to create tests sets to meet Sen et al.'s (2015)'s call to test algorithmic designs against specific communities. Moreover, test sets reflecting the values of specific stakeholder groups should be incorporated into evaluation and reporting toolkits to improve model documentation, building from the work of Gebru et al. (2018) and Mitchell et al. (2019). Non machine learning experts seeking to employ computational models in their work must look to customized test sets to validate potential tools. Using customized test sets, they can validate their approaches against a known set of stakeholders before conducting analyses.

Comparing model test outputs to a chosen population's annotations provides insights on the extent to which a model encodes that population's values and biases. Since test data sets are less costly to create than training data sets due to their smaller size, this approach also helps to address the challenge of gathering training data from difficult-to-reach or very small populations. Such a metric can be used in tandem with traditional accuracy metrics to guide responsible use of computational tools, particularly in the face of challenges brought on by proprietary technologies. In this way, my use of test sets extends work by Mitchell et al. (2019) to provide additional information about appropriate contexts of use for algorithmic models. In addition, my approach to collect survey data about my overall annotator pool could be expanded to collect more data points about just a test data annotator population. Even though fewer annotators are required for test data compared with training data, sufficient data would still be acquired for in-depth analyses that trace social bias.

At the same time, there are practical challenges to turning to nontraditional annotator groups to provide test data. In addition to access and trust, language barriers, familiarity with survey-like prompts, and participant compensation are serious concerns. Crowd workers are notoriously underpaid and maintaining current payment trends with marginalized groups would be highly unethical. Agreement was mixed among my annotator sample, though the specific reason is unclear. One possibility is that, when soliciting annotations, I did not specifically screen for digital skills. Because the panel survey was distributed digitally, participants likely had a minimum baseline digital skill; however, older adults generally have lower digital skill due to less experience with digital technology compared with their younger counterparts (Eshet-Alkalai & Chajut, 2010). Although crowd workers are not a relatively diverse population, their high skill in completing data tasks accurately and efficiently suggests that their agreement is less likely to be influenced by task novelty or fatigue. Collecting data from a stakeholder population may provide relatively little benefit if

inter-annotator agreement is egregiously low. However, there is opportunity to redesign or modify task design to be more accessible to non crowd worker populations. Importantly, low agreement may be caused by low quality data examples, or differing annotator viewpoints. As an example of issues related to data quality, unfamiliar linguistic idiosyncrasies used on Twitter or a lack of social context in tweets may make sentiment annotation difficult for annotators. Considering differing viewpoints, social and cultural context may significantly shape an individual's positive and negative associations with aging-related content, making disagreement an artifact of social experience.

### 5.3.1. Stakeholder Biases

One way of framing the age bias I observed in my analyses is as a misalignment between the values and perspectives on aging that algorithms learned and the values and perspectives on aging that many older adults possess. The older adult bloggers from whom I sampled test inputs expressed many points of view contrary to mainstream stereotypes about aging. However, older adults are underrepresented on the social media sites from which many data sets are scraped, including Twitter, which is the source of Sentiment140. This, coupled with the fact that older adults are underrepresented on crowd sourcing platforms from which training data annotations are often solicited, means that older adults are not the ones defining or influencing how older age is represented in data. In other words, applying a sentiment model which is built from training or annotation data representing a worldview that does not align with that of research subjects, means that, a divergent worldview is being employed that may produce error. The specific biases of research subjects, whether discriminatory or not, are necessary to understand in order to accurately study theirs behavior and attitudes.

Given that algorithms and computational systems will encode the values of the people represented in the underlying data a subsequent and critical question is, then, do the values and biases in our data mirror the values and biases in our application context? The *Age-Related* test set I created took a first step toward answering this question. In the context of studying and measuring human behavior, aligning–or at least identifying– the biases encoded in measurements in relation to the biases possessed by our subjects is important for understanding how and where algorithms err.

CHAPTER 6

# Identifying and Evaluating Unknown Social Biases

In the previous chapter I discussed the importance of unknown associations with age that may be erased through the deletion of training data. Alternatively, preserving both known and unknown associations with age by soliciting input from older adults produced a biased model. Given the trade-offs of each approach, understanding the application context is important for identifying the extent to which known model biases are problematic to apply, as well as for identifying critical associations that a model may not be properly encode. Next I assess whether the social biases encoded in a chosen algorithm mirror the social biases represented in the application context. Up to this point, my investigations have focused on age bias, which is well-defined and emerged readily in the context I chose to investigate. Similarly, many investigations of social bias in algorithms rely on a method of isolating one or several known forms of bias and measuring them in the outputs of algorithmic technologies. Helpful computational tools have been developed through this strategy, such as Aequitas, a tool for auditing algorithmic fairness created by researchers at the University of Chicago (Saleiro et al., 2018). Aequitas analyzes data sets, their annotations, and their identified errors to determine whether algorithmic outputs comply with given fairness criteria. The tool does not specifically identify *social* biases per se, but it can importantly inform algorithm designers if algorithmic outputs are unfairly distributed with respect to input classes, such as race, gender, or age.

This technique, however, relies on sufficient knowledge of distinct forms of social bias before they can be measured or tested. For example, in Studies 1 and 2 had to be familiar with age bias *and* know to test for it. Social biases that researchers test for must be "known unknowns"– or risks that researchers are aware of. But challenges remain for contending with social biases that 1) researchers have not identified or 2) researchers may generally be aware of but ignorant of the fact that they pose a threat to the validity of analyses at hand. How do researchers identify these *unknown* unknowns? At that, even if researchers are aware of potentially relevant social biases, they must be readily measurable in order to be tested and become "known", such as in my isolation of "old" and "young" as age referents. In this chapter I attend to unknown unknowns and directly engage stakeholders to do so.

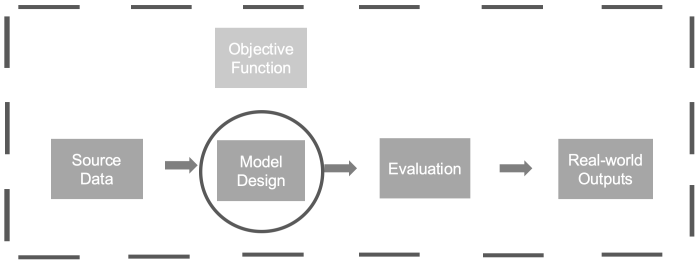


Figure 6.1. Study 3 focuses on the evaluation of the variables defined and considered in a walkability algorithm. This process ultimately raises questions about a number of other stages in the model creation process.

Current work in artificial intelligence has developed computational techniques to address unknown unknowns, such as Lakkaraju et al. (2017); however, I turn to qualitative field methods to understand biases in context as well as to surface biases that may be difficult to measure or quantify. By evaluating an algorithmic tool against the perspectives and values of stakeholders, I identify a slate of social biases that are relevant to the tool's various applications. I invoke Value Sensitive Design (Friedman et al., 2002), which emphasizes values exhibited by humans and computational systems, as a framework for understanding the roots of social bias in systems. Because values shape how both humans and machines prioritize different kinds of data, information, and desired outcomes, I use values as a way to understand human social biases in comparison to social biases in algorithms. In this work I think about social bias as one possible result of values encoded into algorithms and algorithmic technologies.

In the next study, I focus on evaluating model design and the model objective function– that is, the variables defined and optimized in an underlying algorithm. As the work will demonstrate, this evaluation ultimately raises important questions about other decisions in the model creation pipeline (see Figure 6.1). Notably, in Study 3 I depart from my analyses of sentiment models and instead evaluate The Walk Score– a walkability metric used as a proxy for neighborhood livability and resident quality of life1. This intentional shift enables me to more easily engage stakeholders to probe unknown unknowns. In addition to exploring social bias, I consult directly with stakeholders as an additional avenue to incorporate stakeholder perspective and representation. In particular, walkability proves more intuitive to discuss with a variety of people compared to positive and negative sentiment. I turn to semi-structured interviews, which allow the freedom of follow-up questions and exploration of stakeholder biases. This stands in contrast to the surveys and annotation prompts I distributed to annotators Study 2, which inherently rely on a set of pre-defined response choices and limit the range of perspective respondents are able to provide.

1This work was originally published by Diaz and Diakopoulos (2019) at the ACM Conference on Computer Supported Cooperative Work

Specific definitions of walkability may differ from individual to individual or on a group level. In this study, I focus on the ways in which walkability as a concept is defined quantitatively in the Walk Score algorithm and subsequent connections to social bias. By investigating the initial assumptions and definitions used to create the Walk Score, I attend to issues of social bias rooted in a potential mismatch between the values encoded into the design objective of the algorithm (e.g. choices of features and assessments of suitability) and the values of end-users (Kraemer et al., 2010). As such, I specifically take a qualitative field approach to parse these definitions and the implications they may have. Importantly, Nissenbaum and Friedman highlight how shifts in context of use can produce *emergent bias* (Nissenbaum, 2001), and Selbst et al. (2019) highlight a parallel issue in machine learning fairness that they label the 'portability trap'. The definitions and intended application contexts chosen in the initial stage of algorithm design have direct implications for the validity of the resulting model in differing contexts of use. It is for this exact reason that Mitchell et al. (2019) developed a framework for reporting important details about model design and use. Building from their work, qualitative insights can ground model reporting in a sociotechnical context.

### 6.1. Study 3: The Walk Score

The Walk Score is a patented algorithm designed to evaluate walkability ("Walk Score Professional," 2019). Since its introduction in 2007, the Walk Score has garnered attention from professionals and researchers in real estate, urban planning, preventative medicine, and public health as a way of assessing not only walkability, but also the livability of geographic areas. Commercially, the Walk Score is a prominent feature of real estate websites to promote attractive residences to potential renters or buyers; however, researchers in health and medicine have also made extensive use of the Walk Score to study connections between neighborhood walkability, physical activity, and health outcomes. Less studied is the extent to which the Walk Score captures end users' definitions of walkability as well as the range of user biases that influence walking behavior. This research takes aim at assessing (1) alignments and misalignments between users' biases and the biases designed into algorithmic measures of walkability used by researchers and practitioners, and (2) how these alignments and misalignments influence the suitability of algorithmic tools such as the Walk Score in different research contexts. Although I focus my investigation on the Walk Score, this work builds on my examinations of age bias in sentiment analysis and speaks to the need for designers and engineers to evaluate algorithmic tools against specific stakeholder groups and their experiences. I take the Walk Score as just one example of algorithmic tools that aim to measure subjective experience. In undertaking this analysis I explore what the Walk Score does and does not measure about residents' lived experience and its connections to issues of algorithmic bias and transparency.

Though seemingly intuitive, walkability is not a monolithic concept, which means walkability scores are tricky to both generate and interpret systematically. In an investigation of urban planning and mobility, Michael Prescott highlights the issue that walkability literature often ignores disability in discussions of what makes environments walkable (Prescott, 2014). This means, for example, that individuals using wheelchairs cannot rely on common walkability measures to indicate potential barriers to mobility, such as steep inclines or high street curbs. In addition, the Walk Score, specifically, has been criticized for failing to account for typical weather and even whether or not a street has sidewalks, factors that intuitively and significantly impact the likelihood of residents choosing to walk to nearby destinations (de Cambra, 2012). What constitutes suitable walking weather or whether a sidewalk exists can be fairly easily agreed upon; however, the question remains regarding how users differently prioritize aspects of walkability, such as scenery, proximity to various amenities, and street noise. How do individuals conceptualize walkability differently and how might these conceptualizations differ from definitions and biases about walkability purposefully designed into the Walk Score algorithm?

The Walk Score is just one example of algorithmically-produced information that individuals must contend with on a daily basis. As researchers in Human-Computer Interaction seek to understand how best to meet the needs of a variety of end users and appropriately use computational tools, identifying how users differently value information and how information is used by computational algorithms are critically important. The suitability of applying a given computational tool depends, in part, on end users' ability to make appropriate interpretations from the output. Herein lies a tension between generating outputs from easily obtainable data and paying a higher cost for data that may better capture the objective function. In the case of walkability, some factors may be easily quantifiable, such as the number of cafes in a given area, whereas others may be more difficult or even impossible due to subjectivity, such as the neighborhood beauty or the friendliness of neighbors.

Through semi-structured interviews with residents in a large Midwestern U.S. city, I found that respondents drew significant connections between walkability and their social experiences. While the Walk Score encoded participants' general values around walkability, participants indicated that the Walk Score did not account for some significant factors. In particular participants assessed walkability as interwoven with subjective phenomena such as their sense of community, and their sense of safety in the neighborhood. This does not, however, mean that the Walk Score does not provide valuable insight or should not be used in analyses involving walkability. Rather, I call on algorithm designers and researchers to carefully consider how algorithmic metrics such as the Walk Score do or do not serve their specific

goals with respect to communities it is intended to measure and support. For analysts and end users transparency plays an important role in supporting appropriate use and interpretation of algorithmic outputs. I also discuss implications for the ways in which researchers and practitioners use quantitatively measured data produced by tools such as the Walk Score to study phenomena that are significantly influenced by subjective experience. Specifically, I highlight walk equity to describe *whose* walkability is ultimately described by the Walk Score and for whom the algorithmic metric is optimized. This framing can be applied to any algorithmic tool or metric used in analyses of human behavior or experience.

### 6.1.1. The Walk Score in the Real World

The Walk Score works by leveraging the Google Maps API to measure walking distance to amenities in nine different amenity categories, each of which is differently weighted in importance in accordance with published research on walkability (Score, 2014). The Walk Score's amenity categories are *Grocery*, *Bars and Restaurants*, *Retail Shopping*, *Co ee Shops*, *Banks*, *Parks*, *Schools*, *Books*, and *Entertainment*. Generally, only one establishment within each amenity category is counted toward the Walk Score; however, for amenities where increased choice is determined to be important (i.e., *Bars and Restaurants*, *Shopping*, and *Co ee*), multiple amenities are counted, with diminishing value for each additional establishment within that category. The points awarded for a given amenity are a function of these weights as well as a distance decay function for amenities further than .25 miles from a given address. Amenities further than 1.5 miles are not counted. A raw score is comprised of a count of amenities, their weighting, and their distances to a given location. This score is normalized to a scale from 0 to 100. In addition, a score can receive a penalty if the geographic region has long street blocks or low intersection density, which are considered less pedestrian friendly.

In addition to including apartment search tools on its own website, The Walk Score is featured on a number of real estate websites and has an API aimed at real estate professionals and web developers seeking to integrate walkability information into their websites ("Walk Score Professional," 2019). For real estate professionals, the Walk Score is intended as a tool to help advertise the appeal and value of property listings to potential renters and buyers.

Beyond real estate professionals, Redfin, the company that owns the Walk Score, also targets the free-to-use Walk Score API to researchers and analysts in urban planning, government, public health, and finance ("Walk Score Professional," 2019). Work in public health and urban planning is based on an assertion that physical activity increases alongside neighborhood walkability. Indeed, Hirsch et al. (2013) analyzed Walk Score data in relation to survey data,

finding that Walk Scores correlated with respondent activity; however, other research has found limited or nonexistent correlations between Walk Score and physical activity (Jones, 2010; Takahashi et al., 2012).

Although a substantial amount of work exists validating the Walk Score for walkability research (Carr et al., 2011; Duncan et al., 2011; Jones, 2010), this work has been quantitative and largely seeks to validate Walk Score data against other, similar quantitative measurements such as household density or street intersection density. Manaugh and El-Geneidy (2011) found variation in the extent to which the Walk Score correlates with walking behavior based on individual and household characteristics as well as socio-demographic characteristics, suggesting the need for further probing and qualitative insight into what these variations are and what their roots might be. On the other hand, Hirsch et al. (2013) critique prior work that finds little correlation between Walk Scores and walking behavior, pointing out that the prior work is not generalizable due to selection bias. While previous works may indeed lack generalizability, they provide possible evidence of the limitations of the Walk Score for analyzing particular geographic and socio-demographic groups.

It is here that emergent bias materializes as a particular concern. Emergent bias is directly related to a system's context of use and arises when from a change in the application context compared to the original context in which the system was designed or tested, such as population or cultural values (Nissenbaum, 2001). Given that the Walk Score was originally designed for contexts of apartment renting and home buying, investigating walkability in context is necessary to understand aspects that might specifically impact broad-reaching research in public health and urban development. From a critical algorithms perspective, this body of work emphasizes a need to probe why variations in Walk Score validity might exist and what researchers can reasonably interpret from information produced from computational tools such as the Walk Score when studying populations.

### 6.1.2. Value Sensitive Design

Value Sensitive Design provides an approach to understanding how technical systems work, how their design may or may not serve end-user needs, and how to understand the ways in which human values shape the design of systems (Friedman et al., 2002). Beginning with Friedman et al.'s (2008) general definition of a value as, "what a person or group of people consider important in life," I take aim at understanding what individuals consider important in walking and walkability. I then examine the extent to which these considerations and priorities are reflected in the Walk Score. Building on the foundations of VSD, Shilton et al. (2014), outline dimensions of values in the context of sociotechnical systems. Importantly, they underscore the ways in which values can be expressed by users, by systems that humans

design or use, as well as the interactions between humans and systems. They also respond to critiques of VSD calling for clearer methodological frameworks for investigating values in systems by delineating the types of values various research methods are best poised to investigate. Using Shilton et al.'s (2014) outline of value sources and attributes, I found semi-structured interviews to be an appropriate method both to investigate values that were central and peripheral to individuals as well as to complement quantitative studies of the Walk Score and walkability with qualitative insights. In the present work, I begin with an understanding that systems, in addition to humans, are imbued with values, and the interactions between humans and systems shape and are shaped by these values.

In the case of the Walk Score and its use in quantitative analyses, a focus on values helps to parse the facets of walkability that the Walk Score prioritizes as well as both their relationship to social bias and how those facets align with the walkability definitions and priorities of neighborhood residents. The VSD framework allows us to frame the Walk Score as a tool that expresses particular values around walkability. An understanding of these values and those of the individuals the Walk Score is intended to analyze (i.e. neighborhood residents) can help to highlight which elements of residents' experiences are indeed captured by algorithmic metrics such as the Walk score, which elements are not captured, and the kinds of interpretations that are appropriate to make from algorithmic metrics describing aspects of lived experience. Because the Walk Score is used as a proxy for studying neighborhood residents' quality of life, I study residents' values as a point of comparison. Researchers using the Walk Score are implicitly adopting its definitions and values around walkability in their analyses, underscoring the importance of understanding the limits of applying these definitions and values in different research contexts.

### 6.1.3. Critical Algorithm Studies

While the current study homes in on the Walk Score as a site of investigation, my primary interest is in highlighting challenges in representing subjective experience in algorithmic tools, as well as a need to critically evaluate and communicate in- and out-of-scope applications in response to social bias. Algorithmic systems are being leveraged to produce content and analyses in domains ranging from social media to entertainment to criminal justice. As such, researchers have flocked to understand the operations of these systems from both quantitative and qualitative perspectives. The expansion of automated systems has foregrounded the need to develop and test with efficiency and scalability. At the same time, researchers in psychology, design, and human computer interaction are bringing attention to the experiential qualities of algorithm design and user interactions (Alvarado & Waern, 2018; Dietvorst et al., 2015). From both quantitative and qualitative perspectives, there are challenges regarding how to design, implement, and

| Total Population | 49,416 |
|---|---|
| Women | 51.4% |
| Black or African-American | 24.7% |
| Latino or Hispanic | 21.7% |
| White | 45.1% |
| Asian | 5.7% |
| Foreign Born | 26.1% |
| Below Poverty Level | 24.9% |
| Median Household Income | $39,163 |
| High school graduate or higher | 87.4% |

Table 6.1. Neighborhood characterization of population demographics, income, and education according to data from the 2017 American Community Survey Estimates.

evaluate algorithmic systems with respect to underrepresented communities, many of which have unique needs and are disproportionately vulnerable to adverse effects of algorithmic bias (Eubanks, 2018; O'Neil, 2017).

More broadly speaking, it is important to understand the limits of what algorithmic systems can and cannot capture about human experience. Intelligent systems are increasingly managing products and services aimed at broad, diverse audiences (Schrock, 2018; Vlachokyriakos et al., 2016). In particular, an increasing number of cities are making use of smart technologies and intelligent systems to provide crucial services to city residents. It is important to ensure that intelligent systems be responsive to differing information needs and values of users to build trust so that end users can equitably benefit. This project take a values-based approach to exploring these mismatches and their connections to users' understandings, interpretations, and use of algorithmically-produced information.

## 6.2. Method

In order to investigate neighborhood residents' values around walkability, I conducted 14 semi-structured interviews with residents living within a neighborhood of a large Midwestern U.S. city. The following subsections elaborate my recruitment and study procedures.

### 6.2.1. Participants

Basic participant demographics are shown in Tables 6.2 and 6.3. Recruitment was limited to individuals who had been living in the neighborhood for at least one year and who were 18 years or older in age. Participants ranged in age from 25 to 73 and had a median age of 39. In recruitment, I focused on a single neighborhood (see neighborhood details

in Table 6.1). I chose the neighborhood because of its demographic diversity and out of convenience to the proximity of the researchers. Focusing on a single neighborhood allowed me to compare different perspectives on a single Walk Score profile for a geographic area as well as assess how residents differently value the same neighborhood features. Participants were recruited through flyers posted throughout the neighborhood (e.g. in coffee shops, restaurants, bars, and public bulletin boards) and through a neighborhood community group on Facebook. Participants who met the criteria for the study were interviewed on a first-come first-serve basis over a period of two months in late 2018.

### 6.2.2. Procedure

In the first portion of the semi-structured interview, I informed participants that the interview would focus on their experience living in the neighborhood. Drawing from photo elicitation approaches (Harper, 2002), which Le Dantec et al. (2009) demonstrates to be an effective technique in helping respondents voice their values, participants were asked to identify a location in the neighborhood that they frequent or often pass by. Using Google Street View, I then navigated to the location on Google Maps. Google Street View was used to provide a visual stimulus of the location that served to ground further questions. I asked participants why the location was significant and how the location shaped the their experiences living in and navigating around the neighborhood. The first portion of the interview was intended to elicit the kinds of amenities and access residents valued having in their neighborhood.

In the second portion of the interview, participants were asked their thoughts on the walkability of their neighborhood, what informs their decisions to travel by foot, and what might make the neighborhood more walkable for them. These questions helped participants reflect on and concretize their ideas of walkability before evaluating how the Walk Score defines walkability. Next, they were introduced to the Walk Score and asked to estimate the score for the neighborhood before being shown the Walk Score page. This portion of the interview elicited participants' feelings about the extent to which the score reflects their experience living in and navigating around the neighborhood.

Although Walk Scores can be generated for an individual's home address, to protect participant privacy, participants were simply shown the aggregated Walk Score for the neighborhood as well as a visual heat map of walkability for the entire neighborhood to give a sense of the range of walkability scores. Although the Walk Score website does not

| White | Black | Native | Asian | Latinx |
|---|---|---|---|---|
| 8 | 3 | 2 | 1 | 1 |

Figure 6.2. Participant race and ethnicity.

| Male | Female | Median Age |
|---|---|---|
| 7 | 7 | 39 |

Figure 6.3. Participant gender and age.

explain how a neighborhood's or city's *aggregated* Walk Score is specifically calculated or averaged, the visual heat map of walkability shows an overlay of Walk Score variability in a searched area. I briefly explained how the Walk Score algorithm calculates a score, highlighting the specific criteria that contribute to the score. I asked participants their thoughts on the factors that contribute to a Walk Score, which factors they believed were most important, and if there were any factors they would change or include if they could redesign how the Walk Score is calculated. This line of questioning involved using physical cards that could be arranged, representing each of the Walk Score's factors. Blank index cards were provided to allow participants to include their own factors. This second portion of the interview was intended as a simple design exercise and allowed participants to externalize their conceptions of values around walkability.

In the final portion of the interview, participants were asked to give their thoughts on whether they believed the Walk Score provides useful information to them about walkability in their neighborhood and other neighborhoods, whether they trusted the algorithm to generate an accurate score for themselves, and whether they saw benefits or downsides to assessing walkability algorithmically rather than through other methods (e.g., resident reviews, virtual tours). This final series of questions shed light on residents' trust and perceptions of algorithmic tools and the information they produce.

Throughout the interview, I avoided technical jargon such as "algorithm" so that participants would not feel they lacked sufficient knowledge or expertise to provide their thoughts. In total, each interview lasted between 26 and 61 minutes, with a median length of 40 minutes. Participants were interviewed in a public library and a neighborhood coffee shop and were paid $25 for their participation.

### 6.2.3. Data Analysis

Participant interviews were fully transcribed and analyzed using a grounded theory approach (Strauss & Corbin, 1990). I qualitatively analyzed interview transcriptions using iterative inductive analysis, starting with open coding of key concepts relating to walking behavior and priorities around walkability. After the first five interviews, I adjusted the interview protocol to focus on emerging themes. While completing interviews, I memoed and connected related themes, iteratively collapsing them into our final themes, which I report next.

## 6.3. Findings

Overall, the Walk Score seemed to capture participants' general priorities around having walkable access to a variety of amenities.

### 6.3.1. Common Values

Broadly speaking, the Walk Score incorporated a number of factors that participants considered to be important for walkability. In particular, the Walk Score's attention to the distance and density of various amenities aligned with participants' preference for having access to a variety of amenities within close distance. After viewing the Walk Score of their neighborhood, P1 responded, *"That's absolutely how it matches up with what I was saying. Absolutely. I mean, you can do so much."* P9 echoed this sentiment, saying, *"being able to do things with the kids without putting them in a car was important."* The ability to run errands on foot was important to all participants. Supermarkets, for example, were named as particularly important amenities to have close access to, which aligns with the Walk Score algorithm's heavier weighting of grocery stores. When describing the importance of walkability in choosing a place to live, P3 stated, *"I need to have a place that is close to the [train line] and within walking distance to a grocery store"*

### 6.3.2. Di□ering Values

At the same time, participants' individual values around walkability varied somewhat, and these differences were driven by personal context. When presented with the categories of amenities that the Walk Score algorithm uses to calculate scores, participants differed substantially in the categories they identified as most and least important to them for walkability. One example of this emerged when participants discussed the *School* amenity category. For some participants, particularly those with children, schools were highly valued to have within walking distance. P7, a young parent, shared, *"Well, the schools [are important] because we have to go there Monday to Friday, early in the morning. The school and groceries [are] the two biggest things for me. Those are the two things that I do the most."*

For other participants, schools were viewed positively for other reasons. P2 connected the presence of schools to safety saying, *"If you're in the proximity of many schools hopefully, ideally like you're in a safer spot".* P9, a woman in her sixties, valued schools for the diversity they support in the neighborhood, *"I'm going to put schools because I think schools are part of what keeps the diversity in terms of age."* For others, schools were considered less important, or even a nuisance. P5, P10, and P14 named an ambivalence to the presence of schools in their neighborhood because

of their lack of children, with P14 stating, *"I don't have any children, so schools and whatnot wouldn't [be important]"* and P6 actively complained about schools saying, *"When I'm dog walking I go by [two di erent schools] and, I have to say, if it's just when school is getting out that's really annoying because the kids are just crazy."*

Another example of variation between participants was the distance and time individuals were willing to walk to get around. P2 and P4 indicated that they have no issue walking up to 3 or 4 miles to reach a destination, while P8 expressed frustration with previously having lived a 15-minute walk and less than one mile from the closest train station. P1, who developed a foot-related impairment over her time in the neighborhood, did not specifically say how far she is able or willing to walk, but indicated that she still values walking even if it takes more time,

> *"The thing is, it's just a little harder for me now. But the places are still the same. It just takes longer for me to get there, but I can still walk to 'em."*

### 6.3.3. Missing Factors

Participants also named values and biases that were not well-reflected in the Walk Score algorithm's design. This included both individual categories or subcategories of amenities that participants valued, such as transit stops or places of worship, as well as difficult-to-measure elements of day-to-day experiences, such as aesthetic beauty or sense of community.

All participants named the importance and convenience of having a walkable transit stop nearby; however, the Walk Score does not take into account transit accessibility or proximity. All but two participants used a combination of walking and transit as their primary mode of transportation day-to-day, which included completing errands, commuting to work, as well as leisure walks. P4 described their daily routine saying, *"[I walk] almost daily. It's back and forth this way or over to [the train station] and the train."* Redfin (the real estate brokerage that owns the Walk Score) produces a separate Transit Score which, similar to the Walk Score, rates geographic areas, "based on distance and type of nearby transit" ("Walk Score Professional," 2019). The Walk Score and Transit Score exist as distinct metrics. However, participants discussed transit access as an integral component of walkability. P12 described previously living in an area with worse train access compared to their current residence, *"But now we can walk to the train. And so that was something that we were also looking at, like access to specifically the train."* P8 echoed this complaint about train access from a previous residence in comparison to their current home saying, *"because it was a 15 minute walk to the station, to me it was too much."* P3, who does not have a car, described their priorities when seeking a new place to

live, saying, “Transit was a pretty big deal. Personally, I prefer being near the train.” These participants underscored the importance of walkable transit and tended to do so in the context of housing searches, which Redfin markets as a primary context of use for the Walk Score.

Among other amenities missing from the Walk Score algorithm that participants valued were places of worship. P9 highlighted that churches were not factored into the walk score, *”I like being where there’s churches. I mean, I’m not much of a churchgoer. But I do go occasionally.”* P12 and P13, a young Muslim couple, highlighted the importance of having walking access to a mosque and lamented giving up walkable mosque access for other conveniences when moving to their current residence. *”There’s a lot more mosques in the area in general. But yeah, at least where we live in [this neighborhood]. So I would say that we sacrifice [proximity to a mosque] to be closer to like these other things.”* P12 went on to explain a desire for a walkable hair salon and gym in the neighborhood.

### 6.3.4. Subjective Factors

There was a set of subjective features related to social identity that the Walk Score algorithm did not, or could not account for. For values not reflected in the Walk Score, participants often prioritized nebulous or difficult-to-quantify aspects of walkability. Among these were the aesthetics of the neighborhood, personal sense of safety, a sense of community or friendliness in the area, and even neighborhood affordability. These factors interact with walkability in complicated ways and failing to account for them could mean that, at scale, accuracy for particular social groups is compromised.

### 6.3.5. Neighborhood Aesthetics

For participants, the aesthetic beauty, including foliage and the nearby lakefront provided both incentive to go outdoors as well as an improved experience when walking outdoors to accomplish other tasks. P5 highlighted that the natural beauty of the neighborhood is, “almost kind of sort of medicinal, you know, like, it’s really beneficial to my soul.” P1 similarly described the impact of the neighborhood’s beauty on them,

> *”On a nice day, it’s nice. After looking at everything- it’s so much to look at, number one. So that makes the walk go even quicker, you don’t even know, like Oh, I’m right here. So I really enjoy that.”*

This finding echoes quantitative work in urban planning that highlights connections between the design of urban streetscapes and individuals' perceptions of comfort and safety (Harvey et al., 2015), as well as statistical analyses in geography literature linking visual landscapes to perceived walkability (Bereitschaft, 2019). In addition to the natural beauty that other participants highlighted, P11 explained their appreciation for the varied aesthetics that reflect the neighborhood's diversity,

> *"one of the things that I like, like, you walk down [the street] and it's not trying to appear to be like bougie in any way, like, you walk down and it's just like, a Chinese restaurant looks like a Chinese restaurant. A Mexican restaurant looks like a Mexican [restaurant]."*

P11 appreciated that restaurants in the neighborhood maintained a local and *"authentic"* display rather than a more mainstream aesthetic that might cater to wealthier clientele while simultaneously sacrificing cultural roots. Different aesthetic motivations among neighborhood residents means that, as the neighborhood develops and changes over time, residents' interest in outdoor activity may be differentially impacted.

#### 6.3.6. Sense of Safety

In addition to aesthetics, all participants raised concerns about crime and safety in the neighborhood. As with neighborhood transit, the Walk Score website calculates Crime Scores as distinct from walkability; however, participants drew strong connections between walkability and their sense of safety. I found that sense of safety influenced walking behavior, as has been highlighted in prior quantitative assessments of the Walk Score (Towne et al., 2016) and noted as a site for further investigation (Carr et al., 2010). However, I also found that the extent to which crime impacted participants' sense of safety was directly related to demographic characteristics and participants' beliefs about the targeted nature of certain crimes. Not only was crime generally a factor that influenced comfort and walkability in the neighborhood, but also type of crime was a concern for residents.

P1 pointed out that their sense of safety was impacted by time of day. *"[A destination] can be in close proximity, but if it's late, I still want to call [a car]. I'm not gonna just walk those few blocks. It might be a short distance, but if it's late, I'm not gonna do that."* P5 echoed, *"[Crime] is a factor one has to take in consideration when they go out for a walk in the neighborhood."* However, crime was also an accepted reality of living in an urban environment. P1 went on to say, *"we're in a city and things happen."* P7 and P8 reinforced this point saying,

*"There are times that, you know, you hear the gangs and, you know, there's been like, sometimes you hear the gunshots and stu like that, but you're in a city, you know what I mean, so..." -P7*

*"My friends are horrified that I moved to [the neighborhood]. And my response has been, well, I could be hit by a bus tomorrow. So being shot or mugged doesn't bother me like this is part of living in a city is the price you pay to be around lots of people and to have kind of concentrated culture like, if you're going to be living around more people, the chances are, you're going to be living near more bad people so I'm fine with that." -P8*

Participants also highlighted the importance of the *type* and *perceived target* of crime that occurred was of concern. Although they indicated relatively little concern about neighborhood crime, P8, noted that they cared specifically about hate crimes. This concern was rooted in their queer identity and their ability to walk to one of their jobs as a drag queen at a local gay bar.

*"Crime does not bother me as a measure of where I live with one exception, and that is hate crimes...for me, there's a di erence between being burgled or being robbed at gunpoint which feels very opportunistic versus a hate crime which is more personalized and someone once said somewhere I, I forget where I read it, but to live as a queer person, a gay person, whatever is to sacrifice your personal safety in favor of your personal happiness and I feel like I can be whoever I want to be around here." -P8*

To P8, freedom of gender expression and sexual identity, particularly as a person that could be easily identified by others on the street as queer, provided a sense of relative safety. Similarly, when describing two recent murders that occurred in the neighborhood, one of which was a suspected hate crime, P2 stated, *"the hate crime does appear [to be targeted] like that's the scarier one."* Much news about crime in the neighborhood was driven by gang activity, which several participants understood was very targeted. Although they were generally concerned about crime, P2 and P9 recognized the targeted natured of violent crime, noting,

*"In actuality, like, as a single white female I'm probably safer than most because they really don't want that type of attention, to have a crime against a white woman" -P2*

*"I hate the idea that kids in the city are growing up feeling like they're, they're involved, you know, shooting and getting shot at, I hate it. But it didn't make me feel more scared as an individual. Because I thought, 'they're not shooting at me.' So it doesn't make me scared to be out on the street." -P9*

P9 also noted that she had seen crime in neighborhood and city over the course of many years. Rather than grow increasingly wary, she had come understand patterns in crime and navigate her life around it. While the severity or violence of criminal activity did not go unnoticed by participants, residents' perceived risk of danger appeared to play a more significant role in their walking behavior and feelings of safety. This was apparent for older participants who noted that, as older adults, they were more susceptible to others who might attempt to threaten them physically.

### 6.3.7. Sense of Community and Diversity

Sense of community also impacted walking experience for many participants. A number of participants indicated that their sense of community influenced their experiences walking, including their sense of safety.

*"You know, I just wanna step out and get some fresh air. But like I was saying, I'm so close to people that I can just go say hi to, you know" -P1*

*"I do feel more comfortable here than I ever did really anywhere else just because you do have that sense of community... there was a guy who was kind of, a little bit harrass-y, and he followed me home from the [train line] a few times, and so, it was, like, there was a little bit of a sense of like obviously being nervous, but also there were people around that were taking care of me and making sure I was okay, and making sure I was getting home. And so it's like I just don't feel like I would have gotten that anywhere else." -P3*

*"I honestly don't feel any safer than I did 33 years ago. Because there may have been more gangbangers but there were also more grandmas of said gangbangers. Who I'd be like, 'child, get out of my way, I know your grandma', you know."* -P4

For a number of participants this sense of community and comfort in the neighborhood was directly impacted by the racial and ethnic diversity of the neighborhood. P13, who is an immigrant and who speaks Arabic relayed their excitement at encountering a sign in their native language while walking down the street,

> *They had it written in like multiple languages... and the bottom was Arabic and I'm so happy so yeah so as a new person to the neighbor that's really encouraging, yes really encouraging" -P13*

Similarly, P7 drew a connection between neighborhood diversity, community, and local activities,

> *"I think because there is so many people like it helps us to get out more just because we like the di erent cultural events and things that they've had around here...you have to go downtown for a lot of that, you know, yes. So, for it to be right in the neighborhood... it's nice." -P7*

By the same token, several participants valued neighborhood diversity in relation to the variety of local businesses and markets available. P9 shared, *"I shop at [the market] and I get access to all that wonderful stu , all the all the di erent kinds of food from all over the world."*

> *"Just walking distance. Just step out your house, you'll find something good. And so many di erent cuisines. I love the Belizean place on [street], the Jamaican place on [street], several burritos on [street]" -P1*

While the Walk Score for a given area is boosted by the presence of more than one bar or restaurant, the score is not influenced by the variety of cuisines offered.

### 6.3.8. A□ordability

Community and walkability were also interestingly tied to neighborhood development and affordability. When describing their love of the neighborhood and how it has developed over time, P1 explained,

> *"And I always enjoyed it because, hey, we always had the lake to walk to and now it seems like um, certain areas, it's like private land and you have to kind of go around, because the condos now. A ordable housing is a part of everything too because, hey, you know, I want to stay in my place and be able to... why can't I be able to keep walking to all the groovy spots, you know."*

Affordability was commonly cited as something that originally drew participants to move to the area. Although only one participant tied affordability directly to walkability, nearly all participants referenced the rising cost of living in the neighborhood as a concern. When discussing neighborhood safety, P4 also referenced neighborhood affordability,

*"I also don't like that there's a lot of working people who aren't here anymore because they can't a ord to live here. Again, stability. The one house with the mom that would sit on the front porch after school, makes the neighborhood more stable than the 6 people in the condo who are never home, because they're always somewhere else outside the neighborhood."*

Affordability had significant impact on resident activity, namely through access. The ability to afford rent or home costs in the area means residents can gain and maintain access to transit, parks, and a multitude of walkable amenities. For individuals who could no longer afford to stay, moving meant potentially giving up a number of motivations to get outdoors.

### 6.3.9. Interpretability

Although the Walk Score captured factors of general importance to neighborhood residents, ultimately, its perceived usefulness was mixed among participants. This may have been driven by the fact that participants had a tendency to be unsure about the factors that contributed to a give Walk Score. P6 commented that *"[The Walk Score] doesn't tell the whole story"*, while P5 noted, - *"I guess it would be [helpful], I might use it as a resource, but I wouldn't really rely on it entirely. Only because I know...what I like, usually people don't care for and vice versa."* They went on to say, *"I would definitely visit the area in person over the opinion of others, or the score that they have given because like we said earlier things factor that don't really matter to me like schools, etc."* P2, who had used the Walk Score in a housing search in the past, described setting a threshold for filtering out options based on walkability, *"So I'm looking at like 85 or higher on Walk Score. I haven't ended up moving but like that's the... I feel like I've taken places out of the equation. If it's at least eighty five or above all right."* Although P2 did not entirely trust the Walk Score, they did use it as a way to limit their potential housing choices.

Several other participants responded similarly, considering the Walk Score as a potential tool for directing attention but desiring more information about why walkability scores varied and how neighborhood features factored in. P9, an older adult, pointed out a specific desire to live within walking distance of a place where they could fill prescriptions but was unclear on which amenity categories captured that need. They expressed a desire for more information about the Walk Score's underpinnings, *"Well, I'd be interested to see, you know, the areas where the shading is di erent. Why is that?"*. When observing the Walk Score heatmap P5 questioned why Walk Scores decreased closer to the lakefront,

*"Oh, I guess I guess [street] is dark [green]. But why does it get lighter to the lake? I would go dark all the way to the lake."*

When observing the low Walk Scores surrounding a nature park, P5 continued,

*"It's interesting because that's, that's sort of a nature area. But I guess maybe this this website or people maybe maybe thinking more about like, for practical things in their daily life, like, you know, like getting to the train or getting to a job?"* -P5

*"So to me, 80 is pretty high...I don't really know the nuance or the specifics of why it would get down at all."* -P3

When viewing the Walk Score for their neighborhood along with the Walk Score heatmap, participants tended to rationalize differences in Walk Scores that they observed, often remarking on factors that are not actually incorporated into the algorithm. When observing the heatmap, P9 noted their surprise that the lakefront, one of their favorite spots in the area to walk, had a lower Walk Score, saying, *"so like the lakefront shows here as being less walkable. But I suppose that's because of sand."* After noticing that a particular region had a lower Walk Score than expected, P2 tried to rationalize the discrepancy between their personal approximation and that of the Walk Score, *"I'm assuming that I must have misread that and probably how busy the streets are. Maybe people are a little bit more deterred to look around."*

### 6.4. Walkability in Context

From a values perspective, the Walk Score aligns with a wide range of participant values; however, some important subjective factors of walkability were not reflected. The Walk Score also did not capture a number of factors that were significant to participants on an individual basis. Borrowing Shilton et al.'s (2014) values dimensions framework, the Walk Score captured collective values that are relatively central, or highly salient to participants, but in perhaps its most common context of use (providing localized walkability metrics to individual prospective home buyers and renters), the Score breaks down at the level of capturing some influential, individual values. The Walk Score remained a useful source of information, particularly in the context of renter searches; however, some subjective aspects of walkability were not captured, raising questions about whose walkability the Walk Score measures and which users the Walk Score

is designed for. These questions carry important implications for researchers using algorithmic tools such as the Walk Score to study phenomena that are significantly influenced by subjective experiences.

### 6.4.1. Subjective Factors of Walkability

The Walk Score algorithm's approach to measuring walkability primarily relies on tallying physical establishments, avoiding the complexity of measuring subjective factors of walkability while still producing a usable approximation of walkability for prospective renters and buyers. Gillespie's dimensions of *patterns of inclusion* and *the evaluation of relevance* importantly frame the issue of what data is captured in an algorithmic process as well as how it gets captured. "Patterns of inclusion" refers to which data are included and excluded from algorithmic processes, while "the evaluation of relevance" refers to "the criteria by which algorithms determine what is relevant, how those criteria are obscured from us, and how they enact political choices about appropriate and legitimate knowledge" (Gillespie, 2014). In addition to the issue of which factors are weighed more heavily by the Walk Score (e.g., grocery stores), there is the issue of what does not get weighed at all. Despite their absence from the Walk Score algorithm, subjective factors influence walkability in significant ways. Some of these factors prove difficult to quantify, such as a sense of neighborhood community and sense of safety. Nonetheless, these factors influenced participants' experiences while walking as well as their routes and even whether they would opt to walk at certain times.

Adding to the challenge of operationalizing abstract experiences such as a sense of neighborhood community, is that they differ on both an individual basis as well as in systematic ways across social groups. While generally important, the aesthetics and beauty of the neighborhood differed somewhat in priority from participant to participant based on personal opinions of beauty and desire for leisure walking. On the other hand, participants' sense of safety had some roots in social identity. Participants' sense of safety and sensitivity to matters of safety was influenced by racial and gender identity and how these social identities fit into broader social hierarchies. Three participants named their white racial identity as a factor that mitigated safety concerns while out on the street, either because they were less likely to be a target of crime or because they were less likely to be targeted by a police presence. For example, participants recognized that victims of crime are not equally distributed across social identities. This means that some individuals are more likely to be targeted than others and for particular crimes. Similarly, it was noted that police attention was not equally distributed among individuals out on the street. Conversely, sense of safety was more positively impacted by ethnic diversity for some individuals. One participant in a multicultural family and another participant who is arabic-speaking were particularly positively affected by the racial and ethnic diversity in

the neighborhood. Participants' social identities influenced both the salience of walkability factors as well as whether particular concerns, such as police presence, produced positive or negative implications for walkability.

Redfin does not explicitly indicate why it maintains scores for Crime and for Transit separately from the Walk Score, given the implications of those factors on walkability. However, it's plausible that the separation simplifies calculations and helps isolate subjectivities within familiar categories. For instance, incorporating the separate Transit Score would be difficult because of participants' different willingness to take certain forms of public transit as well as different willingness to walk certain distances to reach it. Some kind of composite score might help produce a more robust, general approximation of walking experiences, but the added complexity would also add to the difficulty of making a Walk Score interpretable. For individuals evaluating potential homes, maintaining separate scores allows end users to use their own internal weighting and priorities in interpreting the group of scores.

While the absence of various factors in the Walk Score algorithm is worth noting, designers of similar algorithms should not necessarily seek to create a laundry list of missing factors to begin including. In addition to the inherent challenges in quantifying qualitative experiences, there is no guarantee that designers will have the capacity to uncover a complete list of missing components. Rather, designers must determine the factors most important to account for or preserve for a given, intended analysis such that end-users are able to accurately interpret algorithmic outputs. For factors that may be difficult to operationalize, such as sense of safety, which differs both individually and systematically, designers might consider signalling that such information is not captured in a specific algorithm design. End-users could then account for this missing feature and make decisions regarding how and whether to use algorithmic outputs.

### 6.4.2. Evaluating Contextual and Definitional Suitability via Qualitative Methods

Ultimately, researchers and designers should place emphasis on understanding the context of application and evaluating an algorithmic metric for that context. For example, a walkability metric that does not thoroughly encode racial differences might still be acceptable for group-level analyses among a relatively homogenous population. Based on the Walk Score algorithm's criteria, there is an implicit person for whom walkability is being measured. This person is relatively young and able-bodied, values the particular amenity categories included in the algorithm, and has a particular distance threshold for walking before opting to choose another mode of travel. Less prioritized were older adults, who expressed concern over mobility, infrastructure accessibility, and interests in amenities that have fallen somewhat out of favor among younger generations, such as churches. However, researchers using the Walk Score have little indication of this assumed person. The Walk Score algorithm is relatively transparent; however, its underlying design details are

not well-organized for end users. While the Walk Score does feature documentation on how walkability is measured and scored, little information is organized or packaged to indicate the contexts and types of research questions the tool is best suited to answer, an issue that Mitchell et al. (2019) begin to address with their framework for model reporting. In the end, researchers' assumptions and interpretations of Walk Score data may be false or incomplete. While the Walk Score encodes a particular definition of walkability, researchers must assess this definition against residents' descriptions of walkability as well as their own definitions and assumptions about walkability that may not align with either the Walk Score's or residents' definitions. Such misalignments raise ethical concerns about the validity and potential bias that might be associated with using the score in varied or emerging contexts (Mittelstadt et al., 2016).

While the Walk Score may be sufficiently validated for measuring walkability across a general U.S. population, failing to account for factors that significantly or disproportionately impact the walking behaviors of population subgroups throws into question the validity of the measure for characterizing those particular groups. For example, participants indicated that sense of safety and relationship to neighborhood crime varies with respect to identity. This is something that is likely important to researchers seeking to describe physical activity across populations. From a public health perspective, this could leave subpopulations out of conversations involving policies or strategies aimed at improving health and well-being.

Investigations of what gets captured by the Walk Score and other similar computational tools are important for understanding what these tools actually measure. Understanding the larger context of walkability is critical for designing computational tools or suites of tools and approaches that avoid producing "average" assessments that may not effectively characterize the distribution of individuals' actual lived experiences.

It is here that I highlight the benefits of qualitative methods to investigate assumptions and definitions operationalized in algorithms in a variety of domains. Many investigations of algorithmic bias involve quantitative evaluation of algorithmic outputs with respect to a specific form of bias, such as age-related bias in sentiment analysis (Dıaz et al., 2018). My approach is one that can be undertaken before the deployment of a system or tool, such that end-users and researchers may be provided with insights about *potential* issues in applying an algorithmic model. Mitchell et al.'s (2019) work proposing *model cards*, which are documents that, "disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information", calls for algorithmic model designers to disclose a variety of information to help end-users avoid using algorithmic tools for analyses for which

they are not optimized and may produce unintended bias (Mitchell et al., 2019). A qualitative investigation such as the one I undertook, can complement the quantitative details that Mitchell et al. (2019) call for.

Whereas a post-hoc analysis of algorithmic bias is important for measuring the magnitude of bias and potential mitigation strategies, such an approach requires knowledge a priori of where to look for bias. The approach I took did not test for a specific instance of bias but helped highlight *potential* sources of bias that could be critical for different kinds of analyses. For example, I did not embark on our research with specific intent to delineate racial or socioeconomic underpinnings of walkability; however, these connections first emerged after only just a few interviews. On the other hand, known or specific assumptions can be tested and validated in qualitative ways. Indeed, I was able to confirm factors that the Walk Score was correct to deem important, such as proximity to shopping and entertainment. Pairing our approach with survey methods could further validate or invalidate assumptions across different geographic regions.

Taking a qualitative and exploratory approach such as mine in the initial design of the Walk Score brings clarity to decision making around whether to modify the algorithm in such a way that it begins to address some of these factors or highlight to researcher end-users that these relationships may also be important to consider. Explicitly flagging potential limitations helps researchers with more domain expertise to determine whether to and how best to interpret outputs from an algorithmic tool. Even outside of research contexts, qualitative insight highlights how approaches to data collection might influence subsequent interpretations of outputs— a central focus of Fox et al.'s (2018) design parody exposing socio-spatial exclusions in data-driven maps.

Because my approach takes aim at values to understand differing definitions and assumptions, it is particularly suited for tools that measure phenomena for which user understandings may vary between populations (e.g., men's and women's differing understandings of harassment online (Duggan, 2017); crowd workers' greater likelihood to identify toxic language as hate speech compared with social justice activists (Waseem, 2016)). This includes domains such as toxic comment moderation, where 'toxicity' is defined in different ways (Aroyo et al., 2019). In the case of toxicity, gaining a clear understanding of how end users conceptualize and define toxic behavior is critical for making sure that an algorithm can identify content that users being served indeed believe to be offensive or rude, rather than content that another population believes to be toxic. An approach emulating ours could first try to establish variations in how toxic behavior is defined among members of the population(s) that will be impacted by adoption of a toxicity-detection algorithm and then compare those definitions to the algorithm's parameters. This could highlight instances and

scenarios of toxic behavior for which an algorithm may perform better or worse. Even before prototyping an algorithm, designers might take a similar approach to determine which concepts related to toxicity to attempt to operationalize. Health tracking is another domain in which definitions of a central concept may differ. Health tracking tools can be built on exercise logs that often have different definitions of 'activity level' and the corresponding activities that contribute to metrics (West et al., 2016). End users may have their own distinct notions of healthy activities and falsely assume that an app is measuring those activities. In this instance, a qualitative approach can help to parse assumptions end users make as well as identify to designers how they are successfully or unsuccessfully communicating the algorithm's notion of healthy behavior. The approach I develop in this chapter specifically helps to make different understandings of such subjective concepts designed into algorithms more salient to algorithm designers.

### 6.4.3. Who is the User?

While it is possible that individual consumers apply an implicit understanding of how experiences influenced by their racial identity differ from others when interpreting a Walk Score, researchers seeking to analyze and compare population groups have limited capacity to do the same.

Addressing how to incorporate subjective factors into a walkability metric, or even whether it is necessary to do so, first requires determining the user. Designing a walkability metric requires asking important questions about *whose* walkability should be captured and who needs to capture it. I recommend that designers of algorithmic tools clearly identify stakeholders (including end-users and impacted groups) to then investigate and understand how those stakeholders will interpret algorithmic outputs. The Walk Score is used in different domains, each with a different end user. For example, creating a walkability metric for a prospective home buyer requires algorithm designers to design in response to preferences that vary from individual to individual. Allowing end users to add or remove amenity categories or adjust their relative importance, would support users in optimizing the algorithm for their personal needs. Indeed, Priedhorsky et al. (2012). found that cycling route recommendation algorithms produced results more similar to user ratings when subjective route preferences were considered (Priedhorsky et al., 2012). Personalization could also work in support of explaining why the algorithm produces a specific score, which Rader et al. (2018) find can support awareness of how an algorithm is influencing outcomes. In this case, allowing users to tweak calculations to reflect the values most central and individual to them, or at least be able to identify the "distance" between those values hard-coded into the Walk Score and their own, would support usability. In addition, subjective factors of walkability may differ from user to user substantially enough such that the effort to measure them may provide little value to the

average user. In the same way that maintaining distinct Walk, Crime, and Transit scores may be easier to interpret than a composite walkability score, users can individually weigh their priorities to mentally adjust a given Walk Score to their personal context.

In this work, however, I highlight the use of the Walk Score in population-level research analyses as a critical site of evaluation. The end user in public health and urban development research applications differs from the end user in real estate applications. Supporting usability for researchers and analysts seeking to make generalizations about behavior for different populations requires a focus on supporting accurate interpretations of data outputs. This means that clearly delineating the aspects of walkability that do and do not get captured by the algorithm is necessary. At the same time, it is imperative to make clear which types of analyses the Walk Score is intended to support and is valid for, as well as which analyses it has not been intended or tested for. Delineating these analyses also helps make clear to the algorithm designers whether to embark on operationalizing complicated subjective experiences. For example, applying the Walk Score in an analysis of a racially homogenous and low-crime neighborhood may not warrant incorporation of a 'sense of safety' factor in the way that a highly racially diverse and high-crime area might.

At the same time, systematic variance in these subjective factors across social and community groups means that researchers must consider patterned behavior among subgroups that may deviate from what is typically observed. Considering again the results of Manaugh and El-Geneidy (2011), which found that Walk Score validity varied according to household characteristics, including sociodemographic variables, some of this result may be related to shared characteristics that were not accounted for among the research subjects. Mitchell et al. (2019) again emerges as a significant framework for model creators a to evaluate and clarify to end users the contexts for which a given algorithmic model is well-suited, as well as contexts for which that same model is not well-suited. For algorithm designers, the framework provides a set of references for evaluating a model against its intended use case. Making explicit the intended scope of use would allow designers to either improve walkability metrics or provide explicit signals to the types of analyses and populations for whom the metric is optimized.

### 6.5. Walk Equity

Focusing narrowly on real estate contexts, the Walk Score misses important factors for particular groups that have implications for the utility of walkability scores in low income and rural areas. Notably, the Walk Score does not take into consideration physical infrastructure or the maintenance of infrastructure. For the participants I interviewed, wintry conditions and the extent to which sidewalks were cleared and maintained were significant factors, particularly

for those who felt less stable on their feet due to age or disability. This aligns with Hirsch et al.'s (2017) finding that sociodemographic variables considering physical mobility mediated the relationship between walkability scores and walkability outcomes for Canadian adults. Identifying this issue, researchers with Project Sidewalk have developed ongoing research using mixed-methods approaches to collect and visualize accessibility information at scale about sidewalks across the United States (Hara & Froehlich, 2015; Li et al., 2018). The failure of the Walk Score, in particular, to account for physical infrastructure has been highlighted in past research (Prescott, 2014) and has implications for lower income and rural communities that may be subject to generally worse infrastructure and municipal neglect.

The state of infrastructure is part of a larger discussion around the extent to which the Walk Score equitably defines and characterizes walkability. A failure to account for poorly maintained sidewalk infrastructure in an otherwise well-rated area means that the walkability described is inequitable– that is, it exists for able-bodied residents with the means to navigate rougher paths, while other residents face additional challenges or are barred from walking access. For individuals for whom physical activity may be particularly important, such as folks who are more sedentary due to physical impairments, walkability as measured by the Walk Score may be artificially high and may not take into account important factors needed for them to have access to outdoor physical activity and walking.

Importantly, the quality of infrastructure is tied to the wealth and resources of a given community. Development may improve an objective measure of walkability, but if it is accompanied by displacement, the walkability of the area is enjoyed by a changed population. From a development perspective, improvements in walkability may not be equitably experienced, including any benefits or increase in activity purported to improve the health and well-being of communities. As remarked by one participant, development in her neighborhood limited public access to lakefront areas and increased rent prices, pushing out long-term residents. As tools already in place to advertise desirable homes, the Walk Score and similar walkability metrics may act as forces that prescribe walkability and drive development. The use of the walkability metric as a pointer to neighborhood development and health stands as an example of what Corbett and Loukissas (2019) highlight as a failure to "recognize the experiences of people undergoing gentrification." Particularly for work involving repeated or longitudinal analyses of geographic areas, identifying who experiences changes in development or walkability over time is absolutely necessary for determining whether those changes are equitable. Enjoyment of existing and subsequent walkability features became a benefit for wealthier residents, while displaced residents were forced to relocate to areas further from transportation and other neighborhood amenities.

This matter of *whose* walkability is captured by the Walk Score only became apparent through participant interviews, highlighting the value that qualitative inquiry can provide in investigating ethical design and uses of algorithmic tools.

### 6.5.1. Conclusion

In this chapter, I draw connections between the values in algorithms, their contexts of use, and the potential for resulting social bias. This work also stands as an exploration of how researchers might qualitatively debug objective functions they wish to encode in algorithmic systems. The use of the Walk Score and other computational tools can often rely on quantifying factors that are convenient or relatively easy to quantify. As a result, however, these tools risk ignoring significant components of lived experiences, particularly the lived experiences of underserved populations that designers should aim to support in the creation of computational tools. In order for algorithmic tools to be used ethically and responsibly, researchers must be diligent in determining what they can appropriately infer about data and the lives they describe. One important way of supporting end users is by stating the envisioned context of use and providing increased transparency. Researchers and analysts must closely consider the design of algorithmic metrics and how these metrics may or may not meet their specific analysis goals. Algorithm designers, in turn, must support researchers and analysts in making these determinations by surfacing design specifications. Without this support, designers of algorithms such as the Walk Score risk inadvertently leading researchers and analysts to draw incorrect conclusions about particular geographic areas or about an unspecified subset of the population.

Importantly, my investigations are built on direct consultation with stakeholders of the algorithmic tool I studied. In addition to providing valuable input regarding potential algorithm accuracy, turning to stakeholders is a key component of equitable algorithm design. I call for researchers and practitioners to incorporate understandings of individuals' lived experiences that are often absent in the design and development of technologies as a question of power. This work responds directly to the issue I identified regarding sentiment analysis wherein older adults initially had no part in influencing computational definitions of older age. A variety of methods can help empower and boost stakeholder participation, building off of existing efforts toward ethical algorithmic design .

CHAPTER 7

# Data & Measurement

In this work, I think about capturing the voices and perspectives of stakeholders and underrepresented communities as a form of data representation. Regardless of approach to addressing social bias, I make a call to engage stakeholders in design and evaluation to help elucidate what does and does not get encoded in data and resulting models. Obtaining input from older adults is an approach to encode their biases in data annotations; however, the provenance of the data they are being asked to annotate must be considered as well. That is, the values of the individuals that produce the original data must be considered. If examples in a data set discuss age in a narrow way, for example, annotations from older adults will not be sufficient to expand representation. In my analyses, I asked older adults to provide their perspective vis-a-vis annotations, but I did not ask them to broaden age representation in data by providing new examples.

Understanding the source of training examples is important because social systems, such as those of race, gender, sexuality, and class deeply influence online interactions and the platforms that generate much of the data used in data sets. For algorithmic systems built using data scraped or obtained through digital platforms, data quality is shaped by who participates on a given platform as well as how they participate. Twitter, for example, is disproportionately young in user base, but this phenomenon is not simple happenstance. That Twitter skews young is not in and of itself an issue; however, its use as a massive database of human behavior may be problematic. For example, health insurance companies are interested in Twitter data to infer individual and community health (Neghaiwi, 2016). Twitter users are not representative of society more broadly and treating them as such may lead to modeling errors. An issue akin to emergent bias materializes wherein data production is incentivized for specific platform goals and used in analyses with markedly different goals. For example, users may be incentivized to share certain data to provide platform managers with a higher-fidelity and more valuable data set for advertisers and that data may also be heavily used out of convenience to make inferences about public health. An issue here is that a dearth of users from particular communities, such as older adults, engaging on a social media site means that their data are not represented. However simply choosing an alternative data source is not sufficient. Research in natural language processing has documented

how language models trained on data from news sources with narrow representation of spoken English poorly classify Black vernacular English and other minority dialects (Blodgett & O'Connor, 2017).

Beyond natural language processing, Winifred Poster outlines how racial surveillance pervades online service matching industries, which are platforms that connect actors in service exchange, such as in buying and selling or sharing economy platforms (Benjamin, 2019). She describes that when product images on buying and selling sites like eBay feature Black or darker skinned hands of sellers, the likelihood of sale decreases. She describes how platforms enable and influence surveillance of users that indirectly perpetuates discriminatory practices among platform participants, including both service providers and end-users. These discriminatory practices become encoded in interaction patterns and user behavior in ways that may not be obvious to researchers and engineers eager to leverage convenient web data. Indeed Thebault-Spieker et al. (2017) provide additional empirical analysis to Poster's articulations, showing how mobile crowd-sourcing markets under-represent low-SES regions in systematic ways. The interacting values informing user participation on platforms and that imbue system infrastructures create a slew of complications for understanding whether communities are equitably represented in data. The main takeaway is that the question of who is represented and encoded in an algorithmic technology is inextricably linked to the question of who is represented and encoded in data. The broader influence of various social "isms" (e.g., racism, sexism, classism) mean that disparities in data representation cannot be remedied solely through additional data collection. Those who do not or cannot participate in various systems or platforms will be left out entirely, and those whose participation is limited may be misrepresented.

### 7.1. Mediated through Metrics

As discussed in Chapter 5, social systems such as racism, sexism, and ageism shape platform participation, the data produced from platforms, as well as annotator biases. This influence is further complicated by processes of quantification. Even while engaging in steps to improve data representation and stakeholder voice, researchers must grapple with the ways that quantification mediates both what can be represented in machine learning data sets as well as how algorithmic outputs are interpreted. I am particularly concerned with how underrepresented perspectives in data are quantified in order to be compatible with algorithms. I am also interested in ways that collaborating with marginalized communities can guide algorithm design and challenge the limitations of quantification.

### 7.1.1. Tensions in Quantification

Early adopters and evangelists characterized the Internet as a new frontier that would liberate web users from the social and discriminatory plagues of offline interaction (Barlow, 1996). These early attitudes parallel views on algorithmic systems that amassing enough data might allow us to encode the nuances of human existence to such a degree that analyses afflicted by human biases might be usurped by computational systems. This logic overlooks many necessary social critiques made by scholars in critical race theory and feminism, among other disciplines in the humanities. It also fails to account for the pervasiveness of systemic social issues and their relationship to the limitations of quantification.

In order to be usable to an algorithm, data must be quantified– a process that both allows large magnitudes of data to be efficiently processed as well as limits the kind of data that can be used. Subsequently, human representation in algorithmic data takes limited form. As put by Espeland and Stevens, "numbers may help us comprehend complicated things we care about, but such comprehension comes at the price of mediation." Namely they cite the ways that quantitative measures, "create or reinforce the categories used to conceive of human beings," citing demographic counts and census work as prime examples. Sociology of quantification unpacks the impacts of numbers and quantification on social reality. Rather than understanding numbers and numerical representations, such as graphs, as objective, a sociology of quantification seeks to understand how numbers create social categories, provide legitimacy to ideas, and carry with them persuasive power. In addition, Espeland and Stevens' sociology of quantification paves a way for understanding the limits to quantification. I draw from their insights to delineate how issues of quantification apply to algorithmic systems as well as how qualitative methods can throw into high relief aspects of social experience that algorithmic systems may struggle to encode.

As an example of real-world limitations to quantifying human experience, Green (2020) points out that, of the considerations judges make in determining prison sentences, only a subset can be readily quantified for use in automated risk assessment. This means that some of the considerations that judges are typically expected to weigh, cannot be evaluated at all by automated systems, calling into question how such assessments should be used in criminal justice, if at all. Hearkening back to the introduction of this dissertation, another example is the narrow representation of gender used in automated body scanning that leads to undue difficulty for trans* and nonbinary individuals scanned by TSA in airports. Whether due to too little data or mislabeled data, an algorithm designed upon a flawed representation of gender produces error for a small, but not insignificant portion of the population. In my third research study, Hirsch et al.'s (2013) claim that the Walk Score cannot be expected to perform well for populations that are "too specific",

is a relevant warning that underrepresented populations stand to be harmed by over-reliance on algorithmic metrics intended for broad use.

In light of this, Espeland and Stevens posit a sociology of quantification as a starting point to developing an "ethics of numbers," or a framework for responsibly producing, using, and interpreting numbers. Likewise, an ethics of algorithmic design must draw on similar understandings of quantification– particularly as they relate to technological innovation and our ability to measure marginal experience in algorithmic systems.

### 7.1.2. Categorization and Authority

A sociology of quantification and an ethics of numbers are particularly important for understanding both how marginal experience is translated into data sets, as well as root causes of algorithmic bias that stands to impact underrepresented groups. Even when underrepresented and marginalized communities are the primary focus of research or are otherwise represented in data, their representation is mediated through numbers in complicated ways. For example, categories used to make social groups measurable, such as racial categories, provide interpretability to data that might otherwise be impossible to understand in aggregate. However, racial categories, for example, have notably been unstable historically (Buolamwini & Gebru, 2018; Loveman & Muniz, 2007). The instability of racial and ethnic categories means that race and ethnicity as recorded in one data set may not be comparable with race and ethnicity in another data set. From a historical perspective, "Hispanic" and "Latino" did not become categories in the U.S. census until 1980 and 2000, respectively (Cohn, 2019), meaning adults who today select "Hispanic" or "Latino" had to previously select from the other categories listed. It is necessary to understand the social and cultural limits of specific categorizations so that systems built on them make sense to users and stakeholders. Racial paradigms differ across cultures and may not be intuitive or interpretable outside of their origins. Even when using the same racial and ethnic paradigm, the race of an individual in a single assessment may change across official government records depending on whether race was recorded by the individual, based on phenotypic observation of a criminal justice worker, or copied from another state record (Hanna et al., 2020). From an algorithmic perspective, ground truth data is necessary for an algorithm to learn how to assess an input. The instability of a racial category means that multiple ground truths across data sets may introduce noise to an otherwise singular answer. It is not immediately obvious how the instability of race might impact a specific system used in an influential domain, such as policing or health. Error is an important concern for any algorithmic system; however, empirical work has shown that much algorithmic error is not randomly distributed. Rather, some systems perform more poorly with respect to certain populations (Buolamwini & Gebru, 2018). Because

data about many marginalized groups are already recorded in problematic, incomplete, or inaccurate ways, potential error must be acknowledged and examined in relation to social groups that stand to be further disadvantaged.

Categorizing large amounts of data is one useful way of making it more easily digestible and opens up the possibility of creating visual graphs and models. In describing the aesthetics and conventions of how numbers are visually represented and communicated, such as in statistical graphs, Espeland and Stevens state that, "the most successful numerical pictures influence the ontology of what they represent. The picture becomes its own subject, replaces, in the comprehension of observers, what it originally was intended to merely depict." In distilling information-rich and large data sets into digestible representations, such as metrics, causal models, or graphical representations, these numerical pictures become interpretable stand ins for that which they represent. Likewise, classifications used as proxies in training data sets can stand in as ground truth for phenomena they are intended to approximate or represent. In reference to measurements involving race, Benthall and Haynes assert that, "the creation of metrics and indicators which are race-like will still be interpreted as race" (Benthall & Haynes, 2019).

Proxies that stand in for more complicated phenomena can become particularly problematic because of the *authority* that Espeland and Stevens name as a key dimension of quantification (Espeland & Stevens, 2008). It is precisely this authority that gives numbers the power to create and legitimize categories. Similarly, in their book Sorting Things Out: Classification and Its Consequences, Bowker and Star (2000) explore "the coconstruction of classification systems with the means for data collection and validation." Mirroring Espeland and Stevens (2008)'s assertion that authority legitimizes categories, they note that classification schema inherently "valorize" certain points of view over others, making classification decisions highly influential over social and moral life. Classifications derive from decisions both practical and political. Bowker and Star (2000) assert that "whatever appears as universal or indeed standard, is the result of negotiations, organizational processes, and conflict;" however, these decisions become invisible, "by design and by habit." That is to say that the authority of quantifications and categorizations is not readily apparent to those using or who are otherwise subject to them.

Moreover, authority does not reinforce categories randomly or arbitrarily. Social, cultural, and political forces influence which categories are created and reinforced and, which are not. In this way, numbers not only reify categories, but also extend the underlying logics and values of social systems such as racism and sexism. I uncovered evidence of this in my evaluation of the Walk Score. Implicit in the Walk Score's measurement of walkability are predefined ways of living and being that do not wholly resonate with the experiences of many people. Nonetheless quantified outputs

from the algorithm carry undue credibility that informs accepted truths about the subjects it describes (Kalthoff, 2005; Knorr-Cetina, 1999).

In the same vein, work on algorithmic authority has highlighted the ways that algorithmic decisions can be viewed with similarly undue credibility– even legally (Lustig & Nardi, 2015; Lustig et al., 2016). In criminal justice, algorithmic risk assessments are used in court as unbiased fact, meaning defendants seeking to dismiss them must argue that an assessment is scientifically invalid or that the assessment uses inaccurate information about the defendant (*State v. Loomis.* 2016). That is, if the individual data points used in training are factual, a risk assessment can be considered scientifically valid, legally-speaking, even if that data in the aggregate is flawed– such as arrest records that reflect racist over-policing of Black and Brown populations (Richardson et al., 2019). In addition, intellectual property protections and judges' unfamiliarity with algorithmic technology make challenging the authority of algorithms an uphill battle (*State v. Loomis.* 2016). In addition to shaping social perceptions, quantification and its authority to fix human representation is legally entrenched. In the context of social bias in algorithms, authority rooted in quantification extends the values and perspectives present in system data.

Moreover, fairness criteria addressing bias are also typically quantitatively defined. Selbst et al. (2019) point to a tension between quantitative definitions of fairness and the fact that fairness criteria must be determined by social context. One reason for determining fairness criteria by social context relates to critiques of fairness by Black feminist scholars. Collins (1998), McKittrick (2006) point out that group fairness criteria often treat social groups as equal or interchangeable. Quantifying the historical context that informs nuanced differences in how, for example, an Asian woman and a Black man living in a white heteropatriarchal society experience oppression differently would prove difficult, problematic, or impossible in an algorithmic system. At the same time, Crenshaw (1990) poignantly notes how multiple aspects of social identity produce experiences unique from any individual aspect of social identity. Indeed, D'Ignazio and Klein (2020) assert how data sets can obscure social and historical context. Similarly Cifor et al. (2019) make explicit calls in their Feminist Data Manifest-no for more ethical data practices, among them a refusal to understand data as "disembodied and thereby dehumanized and departicularized". Quantifications lend themselves to interpretation as objective ground truth and must be intentionally reframed through data collection and data analysis. As a result, researchers in FAccT must make efforts to investigate the social contexts in which models will be applied– as well as the social contexts of marginalized stakeholders– so that they can make determinations about how to enforce fairness. Employing qualitative methods provides complementary information to metrics to guide

further quantitative investigations as well as documentation on responsible use. Issues in representing experiences with social discrimination highlight a need to look beyond metrics and quantified characterizations to address bias.

Without careful consideration, the authority of approximations can reinforce flawed logic or misrepresent social groups. In AI fairness research, Hanna et al. (2020) poignantly note that methodologies focused on fairness in relation to protected classes rarely, if ever, account for the fact that race is an unstable social construction. This means that discussions of fairness with respect to race take for granted that race is immutable across time and geography and will fail to trace system error rooted in this fact. Although proxies are not intended to wholly replace a related phenomenon, their use as ground truth can have insidious consequences. In medical research, race is often named as a risk factor when discussing disparate health outcomes across racial groups in the United States, such as the staggering differences in maternal mortality between Black women and women of other races (for Disease Control, (CDC, et al., 1999). This happens despite the fact that scholars have named a necessity to differentiate race from *racism* and other consequences rooted in social inequality, such as diminished healthcare access, as a risk factor in predicting disparate health outcomes (Parker, 1997). In other words, racial categories are often used as convenient proxies for outcomes rooted in social inequality rather than the biology of race. For translation into an algorithmic process, it may be important to quantify population differences. However the choice to operationalize observed racial difference simply as race reifies long debunked assumptions underlying eugenics that racial groups are inherently and biologically different. In ground truth data, then, proxies risk encoding race itself as the canonical root of social inequality. Likewise, algorithmic outputs derived from that data replicate this problematic definition.

### 7.2. Expanding Stakeholder Inclusion

In addition to identifying who produces source data and annotations and understanding their connections to algorithm stakeholders, I call for researchers and organizations invested in machine learning fairness to draw from the humanities as well as participatory methodologies to broaden input on model design and evaluation. Specifically, they must consider broadened inclusion of marginalized and oppressed stakeholders in data collection and data work to address social bias throughout the technology design and use pipeline. With regard to automated risk assessment, Ben Green points out it is “commonly assumed that, with appropriate technical assurances of fairness, risk assessments that inform bail and sentencing decisions can be tailored into neutral tools to improve the criminal justice system” (Green, 2020). Similar attitudes exist regarding computational tools more broadly. The idea is that, if engineers can produce powerful, yet fair, technologies, social bias falls to the wayside as a major issue. However, algorithm

designers must treat all of their designs as risks for propagating harm and cannot discount the potential for social biases to emerge in deployment or future changes in application. Ruha Benjamin critiques this line of thinking in her book *Captivating Technology*. In discussing predictive policing technologies, she calls into question not just predictive policing technology but also the logics underpinning their adoption and use (Benjamin, 2019). She argues that there are inherent issues of bias, surveillance, and control in policing as a larger construct, which means that any technological tool designed to operate within a policing paradigm is inherently flawed. In other words, the values and logics that policing technologies can express are shaped by and limited to those undergirding a carceral state. Issues of injustice linked to punitive logics cannot be solved by shifts in data collection or model debiasing alone. This stands as an example of issues of abstraction that Selbst et al. (2019) highlight in Fairness ML research. They argue that narrow focus in Fairness ML research on specific constraints on fairness metrics abstract away broader context that is necessary for understanding what constitutes fairness in context. It is illustrative that algorithmic systems in criminal justice have largely been used to predict recidivism and risk, which align with punitive logics, rather than rehabilitation or restorative outcomes. As put originally by activist Khadijah Abdurahman, focusing narrowly on questions of operationalization and fairness can distract from injustice with roots outside of algorithm design.

Designers must collaborate with specific communities to elucidate how best to optimize an algorithm in some way, or identify a need to turn to another metric or design altogether. Selecting a paradigm also has significant implications for obtaining annotations in supervised learning. For example, sentiment is often measured according to a "positive" or "negative" (and sometimes "neutral") paradigm. Given that sentiment is expressed in myriad ways across language and culture, it is fitting that there is no single standard for encoding sentiment. However this raises a challenge for how to decide on an appropriate paradigm. Mohammad (2017) outlines difficulties in obtaining consistent, quality sentiment annotations, in part, because of annotation prompts and cases that are difficult to interpret or that are inherently difficult to map to a "positive", "neutral", or "negative" paradigm. In addition to the challenge of selecting a paradigm that aligns with the types of analyses at hand, there may be difficulty in the interpretability of that paradigm for annotators.

Operationalizing a target concept for a given context of use is complicated; however, validating algorithmic designs might be improved through methods that engage insights from stakeholders. Existing work in HCI and FAccT that engages feminist philosophy and critical race perspectives provides a foundation for considering additional methods of expanding communities' role in defining their representation in data and algorithmic systems. I believe there is a moral imperative to work toward this. My intentional centering of older adult perspectives in Study 2 is a response to Shaowen

Bardzell's call for designers and researchers in HCI to begin design work with the perspective of the "marginal user". In the context of algorithmic systems; however, the marginalized person or group may not be a direct end-user. Older adults, or any other group that stands to be unfairly treated by automated systems, would perhaps be better characterized as marginal stakeholders. Still, their perspectives are important to consider in data. Feminist standpoint theory argues that all knowledge is partial. That is, no one person or group has the privilege of a distant or objective view of the truth (Harding, 2004). As previously noted, we see empirical evidence of this in HCI and machine learning research. Patton et al. (2019) at Columbia University found that when annotating social media interactions between gang members, community members more often annotated posts as aggressive, compared with graduate researchers. They posit that community members looked to different features of social media interactions to indicate aggression compared with graduate students, who were less familiar with the social context in which the social media interactions were taking place. Similarly, Sen et al. (2015) found that crowd workers on Amazon Mechanical Turk produce significantly different annotations on gold standard data sets compared with annotators from other communities.

Feminist standpoint theory and empirical work by Patton et al. (2019) and Sen et al. (2015) provide clear evidence that ground truth is not universal, despite what its terminology might imply. However, feminist standpoint theory delineates that these judgments are still truth– just not the *only* truth. Researchers and engineers in machine learning across all industries and organizations must explicitly acknowledge that ground truth–whether provided by human annotators or inferred from human-produced data– is subjective. In this way, ground truth may perhaps be better understood as collective judgments, even for concepts that are seemingly universal, such as walkability. In Study 2, my sample of older adults made certain judgments of age-related content, but those judgments may differ from a sample of older adults that espouse expressly anti-ageist views. Each group's judgments are a version of truth that aligns with their beliefs and perceptions. Researchers and engineers must identify whose ground truth is being captured in order to understand who a computational model is serving.

Given that different communities make different judgments of ground truth, a critically important concern, then, is which ground truth to use in algorithm design. Standpoint theory tells us that, although knowledge is not objective, this does not mean that all claims to truth should be treated equally. People form knowledge from a "situated, embodied location in the world." Thus, for older adults, lived experience with aging informs their views on youth and older age in ways unique from much younger counterparts who have not experienced aging in the same way. Younger individuals have valid and important perspectives on aging; however, in the context of my sentiment analysis research,

and because social systems in our society generally privilege younger age (Butler, 1980), the perspectives of older adults are particularly important to record. In this way standpoint theory and scholarship on age discrimination help to place an imperative on asking older adults to participate in data production, particularly with respect to understanding age-related content. In addition, their perspectives would be altogether missed in a more typical data collection approach that uses online crowd sourcing. Echoing Sen et al. (2015)'s call to evaluate algorithmic technologies against specific communities, Le Dantec et al. (2009) underscore a need to elicit values local to individual stakeholders in service of designing technologies that are responsive to lived experience.

Beyond ground truth data, creating a model that exhibits anti-ageist values, for example, requires careful selection of data and evaluation methods– particularly within the context of a society that is not anti-ageist in its structures and processes. Perspectives espousing anti-ageist views must be incorporated into the processes of data selection, model design, testing, and policies and practices that shape real-world application. Furthermore, an important area of deeper exploration concerns how best to interpret outputs from models trained on data from anti-ageist sources. An end-user may misinterpret outputs if unfamiliar with the goals and motivations of an anti-ageist agenda. Understanding and identifying perspectives to incorporate that challenge mainstream discriminatory ideologies must come through collaborations with people and ideas in the humanities and social sciences. Recent work in HCI and FAccT has additionally called for integrating perspectives from critical race theory to substantively inform system design and evaluation (Hanna et al., 2020; Ogbonnaya-Ogburu et al., 2020). I extend Ogbonnaya-Ogburu et al. (2020)'s call to both acknowledge the burdens of representation on individuals from marginalized communities, as well as make representation an explicit point of conversation in data work and machine learning. There are burdens for individuals from marginalized communities to represent and speak on behalf of an entire social identity group, which is a particular concern in the present dissertation as I frame the need to boost representation in data. In addition to being careful not to over-rely or frame individual's input as wholly representative of any group, Ogbonnaya-Ogburu et al. (2020) note that making conscious efforts to align social identities between members of the research team and the research subjects they engage with is one way to address the burden. Social research has historically had an extractive relationship with marginalized communities (Reverby, 2009), and data collection in service of building broadly applicable algorithmic technologies stands to continue this tradition if not undertaken in ways that acknowledge power differentials.

Documenting data representation extends Gebru et al.'s (2018) proposed paradigm for more standardized data set descriptions. Reporting the characteristics of training and testing data used in algorithmic systems provides

researchers with important context that quantification may struggle to capture. As I previously highlighted in Study 2, historical and social context remains a challenge for algorithmic learning. Incorporating historical context to inform the behavior of algorithmic technologies further complicates the relationship between social movements and algorithmic technologies. As discussed previously in relation to anti-ageism, social movements produce new language and ways of communicating. This inherently adds a challenging layer to algorithmic language detection and understanding. In addition, social movements fundamentally shift societal understanding or acceptance of social norms and behaviors that many algorithmic technologies are designed to detect and analyze. For example, the #MeToo movement has spurred broad-reaching conversations about inappropriate sexual behavior from men in power, as well as men more generally (Mendes et al., 2018). These conversations directly challenge behaviors that have been historically considered appropriate or even the fault of women. Historical moments invoked by #MeToo and related conversations shift mainstream notions of sexually inappropriate behavior and therefore force reassessments of how algorithmic systems define these concepts. However, societal debate surrounding how normative behavior should be defined inherently complicates and politicizes decisions and definitions operationalized into algorithms. For example, men and women make significantly different assessments of sexual harassment online (Duggan, 2017). Researchers must make efforts to understand the implications of privileging one perspective over another, and do so in concert with experts and stakeholders well acquainted with social systems. An algorithmic definition of what constitutes inappropriately sexual communication will inherently be concordant with some views and discordant with others, further emphasizing a need to validate technological systems in relation to particular social contexts and marginalized perspectives.

### 7.3. Improving Data Collection, Generation, and Analysis

Deploying annotation tasks to older adults is an initial step toward critically engaging stakeholders in the design of algorithmic systems. Typical data annotation methods vary widely. Some feature automatic tweet annotation based on the inclusion of specific emoticons (Go et al., 2009), while others employ trained or expert annotators using criteria for removing inconsistent or inaccurate annotations (Hutto & Gilbert, 2014). Nevertheless, all annotated data sets have limitations (Sen et al., 2015). Even with custom annotated data sets, retraining models requires the means to acquire sufficient data, technical skills to pre-process data, as well as time and computational power. Engaging with older adults directly addresses a very practical issue in data collection. Asking older adults themselves to annotate data reduces the cost and effort associated with creating a usable data set while efficiently making use of existing data. Collecting and annotating sufficient data describing underrepresented or marginalized social groups may be more

difficult than for other groups, which places a premium on creating data sets in efficient ways. Because data for specific or underrepresented communities can be more costly to obtain, existing data sets might be combined with smaller data sets annotated by populations of interest in order to create viable, more representative training data sets. Training data creation might be pursued in other targeted ways in terms of population. In addition to annotating an entire data set, specific stakeholders might be engaged in annotating a particular subset of data, as was done in this work, or be asked to generate original training examples. In the context of sentiment analysis, asking a specific population to both generate and annotate examples might provide the highest quality alignment between data and annotation since individuals are best aware of their intended sentiment. Researchers should continue exploring ways of improving existing data sets while also working toward additional methods of generating training and testing examples from stakeholders.

Given the challenges and complexities of quantification, I view collaboration with marginalized stakeholders as important for identifying which aspects of their experiences algorithmic technologies might fail to encode. My evaluation of the Walk Score calls into question how we should operationalize walkability for different purposes. The stakeholder interviews I conducted in Study 3 demonstrate one approach to directly engaging people in algorithm evaluation. I view this work as an initial foray into experimenting with more inclusive data processes in the creation of predictive algorithmic systems. Broadly understanding intersectional issues that impact data representation and quality can be served by Klein and D'Ignazio's call to "bring the bodies back" to data science practice. In understanding collected data and data that should be but has yet to be collected, they argue that "it is people and their bodies who can tell us what data will improve lives and what data will harm them." Participatory methodologies offer additional opportunities to expand on my explorations and collaborate robustly with stakeholders as well as a tradition that has directly engaged with issues rooted in extractive research practices. Similarly, researchers and designers creating fair AI systems can adapt and employ approaches such as participatory methodologies to critically engage marginalized voices in the processes of defining objective functions, operationalizing variables, and identifying the limits of algorithmic technologies. Ultimately, in addition to empowering individuals and communities, closer collaboration with marginalized voices offers an approach to critical reflection on target concepts we choose to measure, their associated indicators, and how we operationalize them. The process of converting data into forms interpretable by machine learning models places limits on how human experience can be algorithmically represented. I argue that directly engaging stakeholders, especially those who have been historically marginalized is a key way of identifying what aspects of lived experience do and do not become encoded in algorithmic designs.

Conveniently, asking older adults to do this work does not fundamentally change the logistics of data collection or validation; however, it does depend on a significant reframing of ground truth as driven by human values and standpoint. That is, annotation tasks can be deployed to older adults in much the same way that they are currently deployed to any number of crowd workers; however, focusing on older adults explicitly acknowledges the unique importance of representing the values and perspectives on ground truth that older adults possess. Disaggregated analyses, which have been highlighted as a critical methodology in fairness assessments (Hanna et al., 2020) can be supported through approaches like the creation of the *Age-Related* test set. A disaggregated analysis will take into account the way fairness metrics change across different categories isolated in data, such as gender or age, and test sets that are specifically crafted to mirror the perspectives of different social groups can provide richer information on the extent to which a model aligns with the views of a particular social group.

### 7.3.1. Participatory Methodologies

In my work, I have drawn on qualitative methodologies toward exploring algorithmic bias and I believe further exploration of qualitative field methods and participatory methodologies, in particular, is an avenue for answering calls by Bardzell (2010), Hanna et al. (2020), Ogbonnaya-Ogburu et al. (2020) for more socially critical work in HCI and machine learning. In bringing attention to the benefits participatory methodologies can contribute to algorithmic design, I direct my call to HCI researchers for several reasons. First, to highlight HCI's tradition of employing a variety of methods, both quantitative and qualitative. Second, to build on the wealth of participatory work among HCI researchers. And third, to extend recent and exciting calls to apply socially critical methods in HCI. For high-stakes applications, participatory approaches provide a balance to the tech mantra of "fail fast, fail often," which downplays the disproportionate cost of failure to marginalized communities.

Taking ground truth as relative, researchers in HCI and fair machine learning must explore participatory methodologies as a site for designing and evaluating computational technologies in response to community issues that are difficult to measure. In HCI, participatory methods have not been often used to study algorithms. Two notable exceptions are Woodruff et al. (2018), who poignantly employed participatory workshops to explore the impacts of algorithmic social bias on end users' perception of systems, and Zhu et al. (2018), who extended insights from Value Sensitive Design to engage stakeholders in the early design stages of a tool for matching Wikipedia contributors to new projects. Additionally, M. K. Lee et al. (2018) used participatory methods to address policy and implementation issues regarding algorithmic fairness. Participatory methodologies are those that specifically engage end users in the

process of designing them (Schuler & Namioka, 1993). Although, in many cases, marginalized stakeholders may not be end users of a given algorithmic technology, there is an opportunity to draw on participatory methods to evaluate technologies and identify potential sites of unfairness during initial design stages in a structured way. With respect to issues of algorithmic social bias, in particular, participatory methods enable the involvement of the wide range of cross-disciplinary stakeholders including those who are nontechnical (Vines et al., 2013). In addition, participatory approaches provide a scaffolded way of identifying potential sites of bias before they emerge in system deployment.

The methods I employed in this dissertation are not strictly participatory, though they seek to begin leveraging stakeholder knowledge. Involving older adults in annotation work provides a way of collecting their insights; however, this form of involvement does not leverage their knowledge toward design or application decisions. Study 3 moves in the direction of participatory methodologies by directly incorporating a short card-sorting design exercise. The card-sorting exercise was focused on model design and understanding how objective function variables relate to stakeholder values. Design decisions and preferences emerged from conversation between myself and my participants in a way that relied on participant experiences to determine design choices that did or did not reflect individual needs. In contrast, I did not interact with data annotators in a way that provided insights about potential issues in how I was measuring sentiment or how I planned to apply sentiment analysis– particularly insights rooted in their first-hand experience with older age and aging.

Ultimately, participatory methods serve to structure the discovery of unknown unknowns and reintroduce context that Selbst et al. (2019) explain is lost through abstraction, which obscures the social dynamics that surround system use. I extend Selbst et al. (2019) to consider how participatory methodologies can clarify how social context shapes data. For example, in Study 1 I focused on the terms “old” and “young” as explicit references to age. Engaging in participatory activities with older adults provides the opportunity for older adults and researchers to discuss ways in which age discrimination manifests in both language and technology design. The result of such activities are twofold. First, researchers can better isolate the kinds of data inputs that might introduce bias. Second, researchers can identify manifestations of age bias that are difficult or impossible to quantify, such that algorithm documentation can clearly indicate blind spots. At that, participatory methods allow researchers to do so in a way that contends with a potential variety of perspectives and disagreements among marginalized stakeholders (Frauenberger et al., 2019). Acknowledging disparate views among individual stakeholder groups and identifying whose interests are being served

is necessary for what scholars in critical race theory have named as a tendency to treat marginalized others as a monolith (Ladson-Billings & Tate, 2000).

Geographer Joni Seager posits that, "if a topic is not evident in standardized databases, then, in a self-fulfilling cycle, it is assumed to be unimportant" Seager (2015). Participatory methodologies can bring light to that which is not evident in standardized data. One challenge of working with data at scale is that "looking at" data typically requires descriptive statistics, histograms, and other quantified summaries. In other words, understanding the people represented in a large data set happens through depersonalized and de-contextualized means. Participatory methods situate researcher, designer, and stakeholder face-to-face, which allows interpretations of stakeholder experience and representation to be shaped by stakeholders themselves. Participatory methodologies scaffold research and design to challenge assumptions, facilitate reciprocal learning, and encourage polyvocal discussions across and through differences (Muller, 2007). Each of these characteristics can help researchers remain reflective in the challenges of selecting and operationalizing both target concepts and their indicators in algorithmic systems. Muller's claim that the goal of participatory methodologies is, "to learn something that we didn't know we needed to know" reflects my motivation in Study 3 to identify 'unknown unknowns' in algorithm evaluation. In addition, participatory methodologies seek substantive involvement of new stakeholder voices through various methods, including storytelling, which scholars in critical race theory name as a powerful means to allow individuals from marginalized communities to articulate their experiences and perspectives for themselves (Ogbonnaya-Ogburu et al., 2020).

I repeat Bowker and Star's (2000) invocation of Latour (1987), that "reality is that which resists" and I view the gaps between stakeholders' experiences and the ways those experiences are calcified and invisibilized into algorithmic solutions as resistances to problematic quantifications. Surfacing these resistances through participatory methods can help algorithm designers to be sensitive to lessons from social theorists while implicitly authority to values and viewpoints encoded in data.

One common participatory method is the workshop. Workshops can take myriad forms and can incorporate a wide range of activities intended to engage new voices in design. As Muller describes, part of the value of workshops is the way in which participants must negotiate new activities together (Muller, 2007). Often held in community spaces, workshops productively challenge traditional notions of expertise because researchers must operate alongside community participants– frequently in community spaces participants are familiar with. At the same time, community participants are able to voice problems, concerns, and ideas with which they have intimate experience. For racial

minorities and immigrant populations, this means giving voice to the ways in which their racial identities do or do not align with typically used paradigms. Researchers can then make determinations about 1. whether a given system uses a paradigm that is applicable to particular groups, and, if not, whether the misalignment is significant enough to warrant new data collection. In addition, direct conversation with stakeholders means that researchers can gain insight about which options individuals choose in the absence of the "ideal" choice.

A wide range of workshop activities, such as storyboarding, scene re-enactment, and photography enable participants to share their perspectives while maintaining narrative control and generate new ideas alongside researchers. In future workshops, for example, participants are led through activities based around critiquing the current state of some process or phenomenon, imagining an improved or alternative future, and beginning to implement imagined changes (Muller, 2007). The portion of my walkability interviews in Study 3 in which I asked participants to redesign a customized walkability algorithm was directly inspired by future workshop-inspired probes of imagining an improved or alternative future through a short design exercise. Participatory workshops expand on this approach.

**7.3.1.1. Participatory Approaches for Model Design.** The scaffolded nature of participatory methodologies affords researchers invested in Fair ML opportunities to adapt them to various stages of the model creation pipeline. At the initial stages of algorithm development, workshops might focus on how elicit stakeholder perspectives on target concepts in context or perhaps make use of algorithm prototypes of different fidelity to garner feedback. Zhu et al. (2018) presented stakeholders with various algorithm prototypes in order to understand implementation issues, performance accuracy, and algorithm alignment with community values. Similarly, (Martin Jr et al., 2020) propose adapting *Community Based System Dynamics* to algorithm development. CBSD is a formal process involving communities to foster a shared understanding of complex systems (Hovmand, 2014). While CBSD and system dynamics ultimately aim to accurately model complex systems, the collaborative process can be used to identify components that evade quantification. From a sociotechnical perspective, collaborative modeling highlights how algorithmic systems interact with social context and help researchers address the traps of abstraction that Selbst et al. (2019) delineates. In this way, there is an opportunity to bring data scientists and engineers together with community participants to develop low- and high-tech prototypes at later stages of development. Ultimately, applying participatory approaches to problem formation as well as objective function formation–as I did in Study 3– has the opportunity to make algorithm design more inclusive and better suited to support specific communities. I believe that in order to develop my earlier example of an anti-ageist computational

model, participatory methodologies must be integrated into the model creation pipeline to guide design decisions with the views and experiences of an anti-ageist community.

Participatory methods can also be applied to improve the annotation collection. As previously mentioned, the data annotation process is not participatory; however, participatory approaches to objective function formation can also incorporate annotation schema. While technical methods exist to identify noisy data points and unclear annotation categories, insight garnered through participatory approaches could distill alternative annotation schema or additional annotation categories that are clear to annotators and stakeholders. For example, workshops might be used to generate a range of annotation schema that can then be tested for specific applications.

In addition, workshops and other field methods can be used to expand the model testing and evaluation toolkit. Although participatory methodologies have not often been applied toward the creation or evaluation of new algorithms, Woodruff et al. (2018) conducted workshops with low-income people of color to explore the qualitative impacts of social bias in algorithmic technologies, highlighting the impacts of unintended social bias in algorithmic technologies as well as stakeholder expectations and trust in company responses to instances of social bias. Woodruff et al.'s (2018) work stands as an excellent demonstration of assessing difficult-to-quantify impacts of algorithmic social bias in real-world applications. Participatory approaches specifically highlight issues in fairness that are difficult or impossible to measure, and prove to be a powerful complement to quantitative fairness approaches currently employed.

Incorporating participatory methods can also help to challenge current approaches to data processing and analysis that typically rely on small groups of data experts and black-box the versions of truth that inform system performance. Eubanks (2018) importantly notes that machine learning practitioners and data scientists represent a relatively narrow slice of lived experience that limits their ability to understand the nuanced impacts of their designs on underserved communities. Recent and important work has called for algorithmic transparency to shift black-boxed approaches toward reporting that provides information about the data and perspectives underlying algorithms (Gebru et al., 2018; Mitchell et al., 2019). Going a step further, D'Ignazio and Klein poignantly call data scientists to consider how data science processes might benefit from an understanding of data work, "not as a genius-like wizardly undertaking, but rather work that embraces multiple voices and valued different types of expertise at all stages of the process." In other words, an approach that explicitly probes values and input from stakeholders as experts. This call directly echoes the underlying motivation of my present research which seeks to leverage the values and expertise of marginal people as reference for comparison in algorithm evaluation. Another perk to participatory methodologies is a way to get direct,

open-ended input about data collection tasks. Such an approach would help elucidate whether my annotation task, which was modeled after typical crowd work tasks, was interpretable to a non crowd worker population.

D'Ignazio and Klein bring attention to ways in which typical data science practices situate data processing and analysis squarely in the domain of a defined data expert who most often lacks domain expertise relevant to the data's originating context. Moreover, the mystique and genius often attributed to the prototypical tech data expert paint data processing and analysis as domains that cannot benefit from the input and interpretations of community members. Moreover computer science and tech often intimidate nontechnical stakeholders from providing useful insight. Collaborating with marginalized communities not only broadens perspectives on data and representation, but also takes into account communities' differing views on fairness and how a system should perform (Grgic-Hlaca et al., 2018). While it is practical to consider the ways in which prevailing views of tech might intimidate or dissuade community participation, it would be false to conclude that community members and relevant non-technical stakeholders could not play a critical role in data collection, processing, or analysis processes. Community activists have played a significant role in bringing attention to and challenging the Chicago Gang Database (Dudek, 2019; Yousef, 2018) and the Data for Black Lives movement brings together data scientists, researchers, and legal scholars alongside community activists and practitioners toward social justice ("Data 4 Black Lives," n.d.). Inviting community activists and organizations to participate in the design and evaluation of algorithmic systems opens up a direct channel of communication on issues of bias. For example, community leaders and members have sophisticated experience with and ideas about the design of technology and technology policy (Dickinson et al., 2019) and can provide insight on past and present issues that exacerbate inequality.

When executed successfully, participatory methods empower community voice and narrative. From an equity standpoint, a significant component of participatory approaches is that stakeholder voices play a role not just in providing valuable information, but also in driving design goals and decisions. In participatory research engaging older adults in community discussions about health, C. N. Harrington et al. (2019) found that participants were empowered to take ownership over how workshops should progress, make choices over individual health management, and collectively take activist approaches to addressing social and environmental determinants of health. Recognizing and supporting the agency of stakeholders in ideating and making decisions about design solutions afforded participants the ability to respond to their lived experiences with systemic inequality. For algorithm design, consulting with stakeholders cannot end with soliciting annotations or other forms of data from a knowledgeable group. As C. Harrington et al. (2019)

state, stakeholders, "should be considered valuable for their knowledge and lived experience in the same way that we consider domain experts in design." That is, participant stakeholders must be treated as co-owners of design. For example, working with stakeholders in model design must entail, not only identifying issues in problem formation or objective function definition, but also allowing participant stakeholders to prioritize which problems are most pressing to formulate and which choices in objective function definition are most critical.

Successfully executing participatory methodologies to produce equitable designs is complex, and algorithm designers will undoubtedly face many challenges adapting participatory methods. However, algorithm designers must begin shifting their approaches to center underserved stakeholders and stakeholders who stand to be most impacted by emerging computational technologies wherever possible. For some analyses, making determinations about which stakeholders will require the most attention may not be obvious–especially for analyses meant to be undertaken at scale and across many social contexts. At the same time, recognizing intersectional experience means explicitly centering stakeholders who exist on the margins, such as trans* women of color or indigenous communities. These communities are among the most neglected in society and, critically, one's membership to these communities does not preclude one's membership to other social groups, such as older adults or immigrant groups. For this very reason, I call for algorithm designers– particularly those working on large-scale or high-stakes technologies– to consider the impacts of their algorithmic designs on the most underserved communities in society, *even when those communities are not explicitly understood as end-users or direct stakeholders*. Rather than take a reactive approach to failures in algorithm design, researchers and engineers must begin from the standpoint that our work risks further harm to already-marginalized people in society.

Although robust in their application, participatory methods are by no means the only domain from which algorithmic fairness researchers can draw. Asynchronous Remote Communities (ARC) have been successfully deployed in HCI research with marginal communities to engage typically unheard perspectives (Walker et al., 2019). And the Delphi method, while not specifically deployed for understanding marginal perspectives, uses an asynchronous, iterative approach to soliciting insight from stakeholders which boost opportunities for qualitative and quantitative error analyses (Baumer et al., 2017). Ultimately, the data science and machine learning pipelines can be complemented by methodological approaches that expand participation and reframe expertise. At the root of my interest in participatory methodologies is a motivation to develop new technologies in a way that restructure social hierarchies that have historically and continue to harm marginalized communities.

CHAPTER 8

# Conclusion

At the highest level, this work asks how researchers can develop algorithmic technologies that are more equitable. Taking the stance that computational systems necessarily contain biases, I explored social bias throughout the algorithmic training and testing pipeline. Focusing on social bias in both data annotations and data provenance, I have considered different approaches researchers may take to address social bias in training data and model outputs. My work contributes in-depth measurements of social bias, its roots in data, and investigates new approaches to control it. I used scholarship in Human-Computer Interaction and Value Sensitive Design as a bridge for understanding how algorithm performance is shaped by human values and beliefs. Unlike prior empirical work on algorithmic social bias, the approach I take ultimately *treats algorithmic bias as an artifact to shape and evaluate for specific contexts* rather than an artifact to remove completely.

My research approach was shaped by qualitative research that threw into question the performance of computational tools. After systematically studying the outputs of popular sentiment models and word embeddings, I found significant age-related bias. This means that these computational tools had a tendency to rate language and text more negatively when they featured references to older age. The application of these tools, then, stands to reify age discrimination already prominent in society. In an attempt to reduce the age bias I observed, I built a custom sentiment model using the training data from a publicly-available model I had tested for bias. However, I modified the data set, removing a relatively small subset of training data that included references to age. This technique eliminated measurable age bias by preventing the learning algorithm from recognizing *any* associations with age. The approach also resulted in little cost to model accuracy.

Despite the success of my approach in reducing age bias, two limitations motivated me to consider other approaches. First, preventing the learning algorithm from learning *any* associations with age meant that my model also had limited capacity to learn connections between age and non-discriminatory phenomena, such as retirement or childhood. For research and analyses focused on older adults, such an approach would be problematic. Second, eliminating bias is not appropriate for understanding the true range of opinions and attitudes among older adults. For example, studying

older adults' attitudes toward a new product or policy would be prone to misinterpretation whether done using a model that is age biased *or* my custom model.

To follow up, I explored an alternative approach rooted in Value Sensitive Design to investigate and reshape age bias present in data sets. I solicited data annotations from a U.S., nationally-representative panel of older adults. Using these annotations, I first created an *Age-Related* test set that can be used as a litmus test to assess whether a chosen language model accurately reproduces the value judgements of older adults.

I also asked older adults to re-annotate the subset of training data I previously removed from my custom sentiment model (the *Older Adult* model). I specifically turned to older adults to provide annotations because they are relevant experts on aging and comprise a group most at risk of being harmed by systems that reproduce age bias. Along with annotations, annotators provided demographic information and responded to a survey assessing their experiences with aging as well as their attitudes toward older age. In this way, I sought to trace a line between model output bias and annotator bias. Building a custom model from the annotations, I found that my model replicated age bias similar to that of the sentiment models I evaluated. It also performed more poorly on my *Age-Related* test set than did a model built from the original training data. Seemingly, my annotation sample reintroduced age bias that I had previously removed.

Interestingly, a qualitative assessment of the disagreements in the outputs of each model suggested that, while both models exhibit age bias, the *nature* of that age bias is somewhat different. Internalized biases that annotators may possess also calls into question whether additional effort to access a non crowd worker population is worth expending to produce a model that still exhibits age bias. My results suggest that to create a model with reduced age bias or that treats age more equitably, collecting training annotations from a general population of older adults is not sufficient on its own. Consulting with older adults to provide custom training examples or involving them in other components of model design may also be necessary. As a result, algorithm designers must be able to identify the extent to which known model biases are problematic for a given research context. They must also be able to identify critical associations that a model may not be encoding.

My final study took aim at uncovering *potential* biases that might emerge based on factors not encoded in a model's design. I turned to qualitative methods to identify previously unknown social biases in relation to a neighborhood walkability metric. Through participant interviews with city residents, I found that significant subjective aspects of walkability are not reflected in the metric, problematizing use of the metric in population-level research analyses. In addition to uncovering biases unknown a priori, this study builds upon scholarship on algorithmic transparency,

providing an approach to evaluate an algorithm in context to surface and report limitations. Although I evaluated an algorithm that is already in widespread use, my approach can be undertaken at the early stages of algorithm prototyping. This final study stands as a demonstration of using qualitative methods to evaluate objective functions against direct input from algorithm stakeholders.

## 8.1. Contributions

Ultimately, this dissertation contributes to scholarship in Critical Algorithm Studies (CAS), Fairness, Accountability, and Transparency (FAccT), and Human-Computer Interaction (HCI). In addition, the data sets I used in this research and the data collected about annotators is available at https://dataverse.harvard.edu/dataverse/algorithm-age-bias/.

Notably, I do not frame social bias as a phenomenon with roots inherent in model building, evaluation, or application. Rather, social bias exists in society and, therefore, the data we use in model testing and training, as well as the data we ultimately analyze using computational tools. While much work emphasizes issues in algorithmic bias and fairness, oftentimes algorithms and the data that underpin them are conflated. It is here that I return to the recurring motif that computational systems are not and cannot be without bias. As a result, we should not unequivocally develop models toward wholly unbiased performance.

Building from CAS and Value Sensitive Design (VSD), I frame bias as a feature of algorithmic systems that researchers can identify, shape, and report to users and stakeholders. My work extends calls in CAS and VSD to grapple with methods to intentionally center marginalized voices in the algorithm design pipeline. I did this quantitatively through data annotation work, as well as qualitatively through field interviews. This dissertation extends work in CAS that highlights the relationship between model data and social systems by empirically exploring annotation data in relation to both model performance and annotator biases. In HCI, my work builds on values-focused scholarship, particularly values sensitive algorithm design first outlined by Zhu et al. (2018) and contributes quantitative and qualitative approaches to assessing values in algorithms.

Finally, to both HCI and FAccT scholarship my work contributes a series of investigations toward molding algorithmic bias for specific contexts of use. Algorithm designers must be able to report on who their designs best serve. My framing of data annotations as an artifact of the annotator's interpretive lens highlights specific ways to do just this, and I call on engineers and researchers to adopt this framing– especially for algorithm applications involving subjective measurements. This framing underscores how models can learn social bias from human judgments in data and forces more careful consideration of annotator samples. In addition to numerous algorithmic fairness auditing

tools that have emerged in recent years, such as Saleiro et al. (2018), attention to data provenance and the human producers of data can help to identify and contextualize biases.

Algorithm designers in both academia and industry must build from the insights of Gebru et al. (2018) and Mitchell et al. (2019) and document socially-oriented information impacting data sources and annotation work. This means reporting on *who* is annotating data and how annotators are similar to or different from algorithm stakeholders. Because all computational systems feature bias, tracing biases back to source data is critical for understanding how to shape model performance to be more equitable. Fast-paced approaches in industry, in particular, must be supplemented with expanded methods, such as more customized test sets and qualitative approaches, in order to address social biases in systems. Tech companies creating new algorithms and partnering with nonprofit and government organizations to create broad-reaching systems must invest in more robust fairness approaches before systems are deployed.

Researchers focused on fairness and accountability can use targeted annotation approaches to create tests sets imbued with the values and judgments of specific communities. Additionally, test sets that reflect the values of specific stakeholder groups can be used to improve model documentation and detail appropriate applications, building from the work of Gebru et al. (2018) and Mitchell et al. (2019). Similarly, FAccT researchers should more often apply qualitative explorations to round out model reporting and guidance for contexts in which social biases and their implications may be unclear, as well as to highlight further contexts that may need additional testing or evaluation. At the same time, researchers can apply qualitative methods to evaluate problem definitions and objective functions before algorithms are even deployed. Finally, fairness researchers and engineers must continue to explore critical inclusion of marginalized stakeholders to deepen contextual understandings of model fairness and performance. My proposal to borrow participatory principles and methods highlights several avenues for follow up research as well as strategic collaboration opportunities for engineers in this domain.

Looking forward, there is opportunity to build from my work in at least two areas. The first is to probe annotator biases more deeply in relation to model performance and help address the question of *whose ground truth should be prioritized and toward which goals*. This is worth particular exploration for the development of algorithmic systems specifically intended to serve marginalized communities and undo past harms. Another area of further exploration is to develop additional channels through which stakeholders can play a role in algorithm development. My research demonstrates that collecting training annotations, alone, is not a sufficient solution to harmful algorithmic bias. Exploring methods of developing novel training data–not just annotations– and employing participatory methods to

engage stakeholder expertise can help shape the values embedded in computational systems at various stages of the design pipeline. Understanding the biases of those who produce training data as well as generating approaches to broaden stakeholder input can help researchers work toward developing a broader framework for ethical algorithm design rooted in the needs and values of marginalized communities. Such a framework must include processes that consider input from stakeholders beyond data scientists, including community advocates and domain experts. Doing so is necessary, not just toward managing bias broadly, but also toward creating technological systems that combat social injustice.

# Bibliography

Alvarado, O., & Waern, A. (2018). Towards algorithmic experience: Initial efforts for social media contexts, In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM. (Cit. on p. 93).

Ananny, M. (2011). The curious connection between apps for gay men and sex offenders. *The Atlantic*, *14* (cit. on p. 47).

Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions, In *Companion proceedings of the 2019 world wide web conference*, San Francisco, USA, ACM. https://doi.org/10.1145/3308560.3317083. (Cit. on p. 109)

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24 (cit. on p. 73).

Axelson, R. D., Solow, C. M., Ferguson, K. J., & Cohen, M. B. (2010). Assessing implicit gender bias in medical student performance evaluations. *Evaluation & the health professions*, *33*(3), 365–385 (cit. on p. 36).

Baker, P., & Potts, A. (2013). ‘‘why do white people have thin lips?” google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, *10*(2), 187–204 (cit. on pp. 31, 47).

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation, In *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics. (Cit. on p. 58).

Bardzell, S. (2010). Feminist hci: Taking stock and outlining an agenda for design, In *Proceedings of the sigchi conference on human factors in computing systems*. (Cit. on pp. 20, 126).

Barlow, J. P. (1996). Declaration of independence for cyberspace. (Cit. on p. 116).

Baumer, E. P., Xu, X., Chu, C., Guha, S., & Gay, G. K. (2017). When subjects interpret the data: Social media non-use as a case for adapting the delphi method to cscw, In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, Portland, Oregon, USA, ACM. https://doi.org/10.1145/2998181.2998182. (Cit. on pp. 49, 132)

Benjamin, R. (2019). *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life*. Duke University Press. (Cit. on pp. 115, 121).

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning, In *Proceedings of the conference on fairness, accountability, and transparency*. (Cit. on p. 118).

Bereitschaft, B. (2019). Exploring perceptions of creativity and walkability in omaha, ne. *City, Culture and Society*, *17*, 8–19 (cit. on p. 100).

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533 (cit. on p. 12).

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). ''it's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions, In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM. (Cit. on pp. 15, 19).

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? inheritance of bias in algorithmic content moderation, In *International conference on social informatics*. Springer. (Cit. on p. 28).

Blodgett, S. L., & O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (cit. on p. 115).

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings, In *Advances in neural information processing systems*. (Cit. on pp. 36, 77).

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121* (cit. on pp. 31, 45).

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press. (Cit. on pp. 118, 128).

Boyd, D., Levy, K., & Marwick, A. (2014). The networked nature of algorithmic discrimination. *Data and Discrimination: Collected Essays. Open Technology Institute* (cit. on pp. 15, 31).

Braithwaite, V. Et al. (2002). Reducing ageism. *Ageism: Stereotyping and prejudice against older persons*, 311–337 (cit. on p. 80).

Brescoll, V. L. (2016). Leading with their hearts? how gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, *27*(3), 415–428 (cit. on p. 36).

Brewer, R., Morris, M. R., & Piper, A. M. (2016). ‘‘why would anybody do this?” understanding older adults’ motivations and challenges in crowd work, In *Proceedings of the 2016 chi conference on human factors in computing systems*. (Cit. on p. 29).

Brewer, R., & Piper, A. M. (2016). " tell it like it really is" a case of online content creation and sharing among older adult bloggers, In *Proceedings of the 2016 chi conference on human factors in computing systems*. (Cit. on p. 29).

Bucher, T. (2012). Want to be on the top? algorithmic power and the threat of invisibility on facebook. *New media & society*, *14*(7), 1164–1180 (cit. on pp. 15, 20).

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47 (cit. on p. 30).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification, In *Conference on fairness, accountability and transparency*. (Cit. on pp. 84, 117).

Butler, R. N. (1969). Age-ism: Another form of bigotry. *The gerontologist*, *9*(4_Part_1), 243–246 (cit. on pp. 29, 76, 80).

Butler, R. N. (1980). Ageism: A foreword. *Journal of Social Issues*, *36*(2), 8–11 (cit. on p. 123).

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292 (cit. on p. 21).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186 (cit. on pp. 20, 23, 28, 31).

Carr, L. J., Dunsiger, S. I., & Marcus, B. H. (2010). Walk score™ as a global estimate of neighborhood walkability. *American journal of preventive medicine*, *39*(5), 460–463 (cit. on p. 100).

Carr, L. J., Dunsiger, S. I., & Marcus, B. H. (2011). Validation of walk score for estimating access to walkable amenities. *Br J Sports Med*, *45*(14), 1144–1148 (cit. on p. 92).

Cascio, J. (2018). What it’s like to go through tsa screening when you’re trans. https://melmagazine.com/en-us/story/what-its-like-to-go-through-tsa-screening-when-youre-trans. (Cit. on p. 12)

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees, In *Proceedings of the conference on fairness, accountability, and transparency*. ACM. (Cit. on p. 21).

Chen, L., Mislove, A., & Wilson, C. (2015). Peeking beneath the hood of uber, In *Proceedings of the 2015 internet measurement conference*. (Cit. on p. 27).

Cifor, M., Garcia, P., Cowan, T., Rault, J., Sutherland, T., Chan, A., Rode, J., Hoffmann, A., Salehi, N., & Nakamura, L. (2019). Feminist data manifest-no. (Cit. on p. 119).

Cohn, D. (2019). Census history: Counting hispanics. https://www.pewsocialtrends.org/2010/03/03/census-history-counting-hispanics-2/. (Cit. on p. 117)

Collins, P. H. (1998). *Fighting words: Black women and the search for justice* (Vol. 7). U of Minnesota Press. (Cit. on p. 119).

Corbett, E., & Loukissas, Y. (2019). Engaging gentrification as a social justice issue in hci, In *Proceedings of the 2019 chi conference on human factors in computing systems*, Glasgow, Scotland Uk, ACM. https://doi.org/10.1145/3290605.3300510. (Cit. on p. 112)

Crawford, K. (2016). Can an algorithm be agonistic? ten scenes from life in calculated publics. *Science, Technology, & Human Values*, *41*(1), 77–92 (cit. on p. 30).

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, *55*, 93 (cit. on p. 21).

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, *43*, 1241 (cit. on pp. 45, 77, 119).

Data 4 black lives. (n.d.). http://d4bl.org/. (Cit. on p. 131)

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys, In *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics. (Cit. on p. 30).

Davies, M. (2008). The corpus of contemporary american english: 450 million words, 1990-present. (Cit. on p. 38).

de Cambra, P. J. M. (2012). *Pedestrian accessibility and attractiveness indicators for walkability assessment* (Doctoral dissertation). Thesis for the Master Degree (MSc) in Urban Studies and Territorial Management. (Cit. on p. 90).

Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction* (Vol. 20). NYU Press. (Cit. on p. 78).

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). " algorithms ruin everything" # riptwitter, folk theories, and resistance to algorithmic change in social media, In *Proceedings of the 2017 chi conference on human factors in computing systems*. (Cit. on p. 26).

Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes (cit. on pp. 21, 27, 29).

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, *3*(3), 398–415 (cit. on pp. 13, 20).

Diaz, M., & Diakopoulos, N. (2019). Whose walkability? challenges in algorithmically measuring subjective experience. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–22 (cit. on p. 88).

Dickinson, J., Dıaz, M., Le Dantec, C. A., & Erete, S. (2019). ‘‘the cavalry ain’t coming in to save us” supporting capacities and relationships through civic tech. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21 (cit. on p. 131).

Dietrich, D. R. (2013). Avatars of whiteness: Racial expression in video game characters. *Sociological Inquiry*, *83*(1), 82–105 (cit. on p. 78).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114 (cit. on p. 93).

D’Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press. (Cit. on pp. 13, 48, 119).

DiSalvo, C., Clement, A., & Pipek, V. (2012). Communities: Participatory design for, with and by communities, In *Routledge international handbook of participatory design*. Routledge. (Cit. on p. 48).

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification, In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society*, New Orleans, LA, USA, ACM. https://doi.org/10.1145/3278721.3278729. (Cit. on p. 20)

Dıaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing age-related bias in sentiment analysis, In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM. (Cit. on pp. 17, 28, 108).

Dombrowski, L., Harmon, E., & Fox, S. (2016). Social justice-oriented interaction design: Outlining key design strategies and commitments, In *Proceedings of the 2016 acm conference on designing interactive systems*. (Cit. on p. 20).

Dudek, M. (2019). Activists want public hearings on chicago police gang database. Chicago Sun-Times. https://chicago.suntimes.com/city-hall/2019/11/13/20963563/chicago-police-gang-database-activists-public-hearings. (Cit. on p. 131)

Duggan, M. (2017). Online harassment 2017 (cit. on pp. 109, 124).

Duggan, M., & Brenner, J. (2013). *The demographics of social media users, 2012* (Vol. 14). Pew Research Center's Internet & American Life Project Washington, DC. (Cit. on p. 49).

Duncan, D. T., Aldstadt, J., Whalen, J., Melly, S. J., & Gortmaker, S. L. (2011). Validation of walk score® for estimating neighborhood walkability: An analysis of four us metropolitan areas. *International journal of environmental research and public health*, *8*(11), 4160–4179 (cit. on p. 92).

ElSahar, H., & El-Beltagy, S. R. (2014). A fully automated approach for arabic slang lexicon extraction from microblogs, In *International conference on intelligent text processing and computational linguistics*. Springer. (Cit. on p. 76).

Eshet-Alkalai, Y., & Chajut, E. (2010). You can teach old dogs new tricks: The factors that affect changes over time in digital literacy. *Journal of Information Technology Education: Research*, *9*(1), 173–181 (cit. on p. 85).

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First i" like" it, then i hide it: Folk theories of social feeds, In *Proceedings of the 2016 chi conference on human factors in computing systems*. (Cit. on p. 26).

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds, In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. ACM. (Cit. on pp. 15, 26).

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401–436 (cit. on pp. 14, 118).

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. (Cit. on pp. 13, 15, 20, 94, 130).

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89 (cit. on p. 30).

for Disease Control, C., (CDC, P. Et al. (1999). State-specific maternal mortality among black and white women–united states, 1987-1996. *MMWR. Morbidity and mortality weekly report*, *48*(23), 492 (cit. on p. 120).

for Disease Control, C., (CDC, P. Et al. (2003). Trends in aging–united states and worldwide. *MMWR. Morbidity and mortality weekly report*, *52*(6), 101 (cit. on p. 49).

Fox, S. E., Lampe, M., & Rosner, D. K. (2018). Parody in place: Exposing socio-spatial exclusions in data-driven maps with design parody, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, ACM. https://doi.org/10.1145/3173574.3173896. (Cit. on p. 109)

Frauenberger, C., Spiel, K., Scheepmaker, L., & Posch, I. (2019). Nurturing constructive disagreement-agonistic design with neurodiverse children, In *Proceedings of the 2019 chi conference on human factors in computing systems*. (Cit. on p. 127).

Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2018). Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured twitter data. *Social Science Computer Review*, 0894439318788314 (cit. on p. 28).

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning, In *Proceedings of the conference on fairness, accountability, and transparency*. ACM. (Cit. on p. 21).

Friedman, B. (1997). *Human values and the design of computer technology*. Cambridge University Press. (Cit. on p. 20).

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. *The handbook of information and computer ethics*, 69–101 (cit. on p. 92).

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, (02–12) (cit. on pp. 16, 88, 92).

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347 (cit. on pp. 15, 20).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (cit. on pp. 13, 22, 85, 123, 130, 136).

Gillespie, T. (2014). The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, *167* (cit. on pp. 20, 23, 82, 106).

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009 (cit. on pp. 41, 124).

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (cit. on p. 78).

Green, B. (2018). ‘‘fair" risk assessments: A precarious approach for criminal justice reform, In *5th workshop on fairness, accountability, and transparency in machine learning*. (Cit. on p. 22).

Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness, In *Proceedings of the conference on fairness, accountability, and transparency (fat*'20). acm. https://doi. org/10.1145/3351095.3372869*. (Cit. on pp. 116, 120).

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments, In *Proceedings of the conference on fairness, accountability, and transparency*. (Cit. on p. 26).

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, *74*(6), 1464 (cit. on p. 76).

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction, In *Proceedings of the 2018 world wide web conference*. (Cit. on p. 131).

Guo, P. J. (2017). Older adults learning computer programming: Motivations, frustrations, and design opportunities, In *Proceedings of the 2017 chi conference on human factors in computing systems*. (Cit. on p. 29).

Hajian, S., & Domingo-Ferrer, J. (2012). A study on the impact of data anonymization on anti-discrimination, In *2012 ieee 12th international conference on data mining workshops*. IEEE. (Cit. on p. 21).

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, ACM. https://doi.org/10.1145/3173574.3173582. (Cit. on p. 23)

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness, In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. (Cit. on pp. 13, 16, 117, 120, 123, 126).

Hara, K., & Froehlich, J. E. (2015). Characterizing and visualizing physical world accessibility at scale using crowdsourcing, computer vision, and machine learning. *ACM SIGACCESS Accessibility and Computing*, (113), 13–21 (cit. on p. 112).

Harding, S. G. (2004). *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press. (Cit. on p. 122).

Harley, D., & Fitzpatrick, G. (2009). Youtube and intergenerational communication: The case of geriatric1927. *Universal access in the information society*, *8*(1), 5–20 (cit. on p. 29).

Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual studies*, *17*(1), 13–26 (cit. on p. 95).

Harrington, C. N., Borgos-Rodriguez, K., & Piper, A. M. (2019). Engaging low-income african american older adults in health discussions through community-based design workshops, In *Proceedings of the 2019 chi conference on human factors in computing systems*. (Cit. on pp. 81, 131).

Harrington, C., Erete, S., & Piper, A. M. (2019). Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–25 (cit. on p. 131).

Harvey, C., Aultman-Hall, L., Hurley, S. E., & Troy, A. (2015). Effects of skeletal streetscape design on perceived safety. *Landscape and Urban Planning*, *142*, 18–28 (cit. on p. 100).

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories, In *Proceedings of the fourth international conference on communities and technologies*. (Cit. on p. 63).

Hirsch, J. A., Moore, K. A., Evenson, K. R., Rodriguez, D. A., & Roux, A. V. D. (2013). Walk score® and transit score® and walking in the multi-ethnic study of atherosclerosis. *American journal of preventive medicine*, *45*(2), 158–166 (cit. on pp. 91, 92, 116).

Hirsch, J. A., Winters, M., Clarke, P. J., Ste-Marie, N., & McKay, H. A. (2017). The influence of walkability on broader mobility for canadian middle aged and older adults: An examination of walk score™ and the mobility over varied environments scale (moves). *Preventive medicine*, *95*, S60–S67 (cit. on p. 112).

Holdren, J. P., & Lander, E. S. (2016). *Technology and the future of cities* (tech. rep.). Executive Office of the President. (Cit. on p. 12).

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (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*. ACM. (Cit. on p. 19).

Hovmand, P. S. (2014). Group model building and community-based system dynamics process, In *Community based system dynamics*. Springer. (Cit. on p. 129).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews, In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining*. (Cit. on p. 30).

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes, In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*. Association for Computational Linguistics. (Cit. on p. 30).

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the implicit association test to measure age differences in implicit social cognitions. *Psychology and aging*, *17*(3), 482 (cit. on p. 76).

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, In *Eighth international aaai conference on weblogs and social media*. (Cit. on p. 124).

Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., & Sips, R.-J. (2014). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data, In *International semantic web conference*. Springer. (Cit. on p. 66).

Introna, L. D., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The information society*, *16*(3), 169–185 (cit. on pp. 15, 20).

Introna, L., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society*, *2*(2/3), 177–198 (cit. on pp. 15, 20, 47).

Irani, L., Vertesi, J., Dourish, P., Philip, K., & Grinter, R. E. (2010). Postcolonial computing: A lens on design and development, In *Proceedings of the sigchi conference on human factors in computing systems*. (Cit. on p. 20).

Johnson, I., McMahon, C., Schöning, J., & Hecht, B. (2017). The effect of population and "structural" biases on social media-based algorithms: A case study in geolocation inference across the urban-rural spectrum, In *Proceedings of the 2017 chi conference on human factors in computing systems*. (Cit. on p. 27).

Jones, L. I. (2010). *Investigating neighborhood walkability and its association with physical activity levels and body composition of a sample of maryland adolescent girls* (Doctoral dissertation). (Cit. on pp. 18, 92).

Juarez, J. A., & Brown, K. D. (2008). Extracting or empowering? a critique of participatory methods for marginalized populations. *Landscape Journal*, *27*(2), 190–204 (cit. on p. 48).

Kalthoff, H. (2005). Practices of calculation: Economic representations and risk management. *Theory, Culture & Society*, *22*(2), 69–97 (cit. on p. 119).

Kamiran, F., & Calders, T. (2009). Classifying without discriminating, In *2009 2nd international conference on computer, control and communication*. IEEE. (Cit. on p. 21).

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, *35*(3), 613–644 (cit. on p. 21).

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer, In *Joint european conference on machine learning and knowledge discovery in databases*. Springer. (Cit. on p. 22).

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations, In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. (Cit. on p. 31).

Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 88:1–88:22. https://doi.org/10.1145/3274357 (cit. on p. 23)

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29 (cit. on p. 27).

Kizilcec, R. F. (2016). How much information?: Effects of transparency on trust in an algorithmic interface, In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM. (Cit. on p. 83).

Knorr-Cetina, K. (1999). Epistemic cultures: The cultures of knowledge societies. *Cambridge, MA: Harvard* (cit. on p. 119).

Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, *22*(2), 100–109 (cit. on p. 60).

Kraemer, F., van Overveld, K., & Peterson, M. (2010). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260 (cit. on p. 89).

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media, In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. ACM. (Cit. on p. 26).

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct. *Harvard business review*, *92*(5), p32–32 (cit. on p. 29).

Ladson-Billings, G., & Tate, W. F. (2000). Toward a critical race theory of education. *Sociology of Education: Major Themes, edited by Stephen Ball*, 322–342 (cit. on p. 128).

Lahey, J. N. (2010). International comparison of age discrimination laws. *Research on aging*, *32*(6), 679–697 (cit. on p. 29).

Lakkaraju, H., Kamar, E., Caruana, R., & Horvitz, E. (2017). Identifying unknown unknowns in the open world: Representations and policies for guided exploration, In *Thirty-first aaai conference on artificial intelligence*. (Cit. on p. 88).

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, *9* (cit. on p. 84).

Lasher, K. P., & Faulkender, P. J. (1993). Measurement of aging anxiety: Development of the anxiety about aging scale. *The International Journal of Aging and Human Development*, *37*(4), 247–259 (cit. on pp. 61, 62, 76).

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard university press. (Cit. on p. 128).

Lazar, A., Diaz, M., Brewer, R., Kim, C., & Piper, A. M. (2017). Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism, In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, Portland, Oregon, USA, ACM. https://doi.org/10.1145/2998181.2998275. (Cit. on pp. 20, 29, 33, 39, 45, 52, 75, 81)

Le Dantec, C. A. (2016). *Designing publics*. MIT Press. (Cit. on pp. 48, 49).

Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). Values as lived experience: Evolving value sensitive design in support of value discovery, In *Proceedings of the sigchi conference on human factors in computing systems*. ACM. (Cit. on pp. 95, 123).

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division, In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. ACM. (Cit. on p. 22).

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., Noothigattu, R., See, D., Lee, S., Psomas, C.-A., Et al. (2018). Webuildai: Participatory framework for fair and efficient algorithmic governance. *Preprint* (cit. on p. 126).

Levy, B. (2009). Stereotype embodiment: A psychosocial approach to aging. *Current directions in psychological science*, *18*(6), 332–336 (cit. on p. 63).

Li, A., Saha, M., Gupta, A., & Froehlich, J. E. (2018). Interactively modeling and visualizing neighborhood accessibility at scale: An initial study of washington dc, In *Proceedings of the 20th international acm sigaccess conference on computers and accessibility*. ACM. (Cit. on p. 112).

Liao, Q. V., Fu, W.-T., & Strohmaier, M. (2016). # snowden: Understanding biases introduced by behavioral differences of opinion groups on social media, In *Proceedings of the 2016 chi conference on human factors in computing systems*. (Cit. on p. 26).

Loveman, M., & Muniz, J. O. (2007). How puerto rico became white: Boundary dynamics and intercensus racial reclassification. *American Sociological Review*, *72*(6), 915–939 (cit. on p. 117).

Lustig, C., & Nardi, B. (2015). Algorithmic authority: The case of bitcoin, In *2015 48th hawaii international conference on system sciences*. IEEE. (Cit. on p. 119).

Lustig, C., Pine, K., Nardi, B., Irani, L., Lee, M. K., Nafus, D., & Sandvig, C. (2016). Algorithmic authority: The ethics, politics, and economics of algorithms that interpret, decide, and manage, In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems*, San Jose, California, USA, ACM. https://doi.org/10.1145/2851581.2886426. (Cit. on p. 119)

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis, In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics. (Cit. on p. 30).

Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., & Saxena, M. (2018). Analyze, detect and remove gender stereotyping from bollywood movies, In *Conference on fairness, accountability and transparency*. (Cit. on p. 21).

Manaugh, K., & El-Geneidy, A. (2011). Validating walkability indices: How do different households respond to the walkability of their neighborhood? *Transportation research part D: transport and environment*, *16*(4), 309–315 (cit. on pp. 92, 111).

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank (cit. on p. 58).

Martin Jr, D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Isaac, W. S. (2020). Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (cit. on pp. 15, 16, 129).

Matsumoto, K., Akita, K., Keranmu, X., Yoshida, M., & Kita, K. (2014). Extraction japanese slang from weblog data based on script type and stroke count. *Procedia Computer Science*, *35*, 464–473 (cit. on p. 76).

McKittrick, K. (2006). *Demonic grounds: Black women and the cartographies of struggle*. U of Minnesota Press. (Cit. on p. 119).

Mendes, K., Ringrose, J., & Keller, J. (2018). #metoo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, *25*(2), 236–246 (cit. on p. 124).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*. (Cit. on p. 36).

Miller, B., & Record, I. (2017). Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. *New Media & Society*, *19*(12), 1945–1963 (cit. on p. 29).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting, In *Proceedings of the conference on fairness, accountability, and transparency*. ACM. (Cit. on pp. 13, 22, 24, 47, 85, 89, 108, 109, 111, 130, 136).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data Society*, *3*(2) (cit. on pp. 24, 108).

Mohammad, S. M. (2017). Challenges in sentiment analysis, In *A practical guide to sentiment analysis*. Springer. (Cit. on pp. 73, 121).

Muller, M. J. (2007). Participatory design: The third space in hci, In *The human-computer interaction handbook*. CRC press. (Cit. on pp. 128, 129).

Neghaiwi, B. H. (2016). In insurance big data could lower rates for optimistic tweeters. Thomson Reuters. https://www.reuters.com/article/us-insurers-bigdata-consumers-idUSKCN12N05R. (Cit. on p. 114)

Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 120–119 (cit. on pp. 13, 15, 20, 27, 82, 89, 92).

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. (Cit. on pp. 15, 20, 31).

Officer, A., Schneiders, M. L., Wu, D., Nash, P., Thiyagarajan, J. A., & Beard, J. R. (2016). Valuing older people: Time for a global campaign to combat ageism. *Bulletin of the World Health Organization*, *94*(10), 710 (cit. on pp. 29, 76).

Ogbonnaya-Ogburu, I. F., Smith, A. D., To, A., & Toyama, K. (2020). Critical race theory for hci, In *Proceedings of the 2020 chi conference on human factors in computing systems (chi'20). https://doi. org/10.1145/3313831.3376392*. (Cit. on pp. 78, 123, 126, 128).

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books. (Cit. on pp. 13, 94).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135 (cit. on p. 30).

Parker, H. (1997). Beyond ethnic categories: Why racism should be a variable in health services research. *Journal of health services research & policy*, *2*(4), 256–259 (cit. on p. 120).

Pasquale, F. (2015). *The black box society*. Harvard University Press. (Cit. on p. 29).

Patton, D. U., Blandfort, P., Frey, W. R., Gaskell, M. B., & Karaman, S. (2019). Annotating twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators (cit. on pp. 28, 58, 122).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830 (cit. on p. 53).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation, In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*. (Cit. on p. 38).

Posch, L., Bleier, A., Flöck, F., & Strohmaier, M. (2018). Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *arXiv preprint arXiv:1812.05948* (cit. on pp. 49, 58).

Prescott, M. (2014). Using social topography to understand the active mobility networks (amns) of people with disabilities (pwds). UWSpace. http://hdl.handle.net/10012/8250. (Cit. on pp. 14, 90, 112)

Priedhorsky, R., Pitchford, D., Sen, S., & Terveen, L. (2012). Recommending routes in the context of bicycling: Algorithms, evaluation, and the value of personalization, In *Proceedings of the acm 2012 conference on computer supported cooperative work*, Seattle, Washington, USA, ACM. https://doi.org/10.1145/2145204.2145350. (Cit. on p. 110)

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, ACM. https://doi.org/10.1145/3173574.3173677. (Cit. on p. 110)

Reverby, S. M. (2009). *Examining tuskegee: The infamous syphilis study and its legacy*. Univ of North Carolina Press. (Cit. on p. 123).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘‘why should i trust you?” explaining the predictions of any classifier, In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. (Cit. on p. 32).

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, Forthcoming* (cit. on p. 119).

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep* (cit. on p. 49).

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (cit. on pp. 22, 87, 136).

Schrock, A. (2018). Civic tech: Making technology work for people. Rogue Academic Press, California. (Cit. on pp. 23, 94).

Schuler, D., & Namioka, A. (1993). *Participatory design: Principles and practices*. CRC Press. (Cit. on pp. 48, 127).

Score, W. (2014). Walk score methodology. *Accessed April*, *24* (cit. on p. 91).

Seager, J. (2015). *Sex-disaggregated indicators for water assessment, monitoring and reporting*. UNESCO Publishing. (Cit. on p. 128).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems, In *Proceedings of the conference on fairness, accountability, and transparency*. (Cit. on pp. 89, 119, 121, 127, 129).

Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., Russell, J., Wang, Z. K., & Hecht, B. (2015). Turkers, scholars, arafat and peace: Cultural communities and algorithmic gold standards, In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*. ACM. (Cit. on pp. 15, 17, 28, 30, 47, 50, 58, 66, 85, 122–124).

Shilton, K., Koepfler, J. A., & Fleischmann, K. R. (2014). How to see values in social computing: Methods for studying values dimensions, In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing*. ACM. (Cit. on pp. 92, 93, 105).

Smith, A. (2014). Older adults and technology use. pew research center. *Retrieved October*, *2*, 2015 (cit. on p. 45).

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks, In *Proceedings of the 2008 conference on empirical methods in natural language processing*. (Cit. on p. 58).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank, In *Proceedings of the 2013 conference on empirical methods in natural language processing*. (Cit. on p. 31).

Starbird, K., & Palen, L. (2012). (how) will the revolution be retweeted? information diffusion and the 2011 egyptian uprising, In *Proceedings of the acm 2012 conference on computer supported cooperative work*. (Cit. on p. 75).

*State v. loomis*. (Vol. 881). (2016). (Cit. on p. 119).

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Inc. (Cit. on p. 96).

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10–29 (cit. on p. 31).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307 (cit. on p. 30).

Takahashi, P. Y., Baker, M. A., Cha, S., & Targonski, P. V. (2012). A cross-sectional survey of the relationship between walking, biking, and the built environment for adults aged over 70 years. *Risk management and healthcare policy*, *5*, 35 (cit. on p. 92).

Thebault-Spieker, J., Terveen, L., & Hecht, B. (2017). Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Trans. Comput.-Hum. Interact.*, *24*(3), 21:1–21:40. https://doi.org/10.1145/3058499 (cit. on pp. 20, 115)

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias, In *Cvpr 2011*. IEEE. (Cit. on p. 30).

Towne, S. D., Won, J., Lee, S., Ory, M. G., Forjuoh, S. N., Wang, S., & Lee, C. (2016). Using walk score™ and neighborhood perceptions to assess walking among middle-aged and older adults. *Journal of community health*, *41*(5), 977–988 (cit. on p. 100).

Twyman, M., Keegan, B. C., & Shaw, A. (2017). Black lives matter in wikipedia: Collective memory and collaboration around online social movements, In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. (Cit. on p. 75).

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530 (cit. on p. 21).

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making, In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM. (Cit. on p. 23).

Venables, W., & Ripley, B. (2002). Modern applied statistics with s. springer-verlag. *New York* (cit. on p. 33).

Vines, J., Clarke, R., Wright, P., McCarthy, J., & Olivier, P. (2013). Configuring participation: On how we involve people in design, In *Proceedings of the sigchi conference on human factors in computing systems*. (Cit. on p. 127).

Vines, J., Pritchard, G., Wright, P., Olivier, P., & Brittain, K. (2015). An age-old problem: Examining the discourses of ageing in hci and strategies for future research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *22*(1), 2 (cit. on p. 29).

Vlachokyriakos, V., Crivellaro, C., Le Dantec, C. A., Gordon, E., Wright, P., & Olivier, P. (2016). Digital civics: Citizen empowerment with and through technology, In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems*. ACM. (Cit. on pp. 23, 94).

Voida, A., Dombrowski, L., Hayes, G. R., & Mazmanian, M. (2014). Shared values/conflicting logics: Working around e-government systems, In *Proceedings of the sigchi conference on human factors in computing systems*. ACM. (Cit. on p. 16).

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in wikipedia. *EPJ Data Science*, *5*(1), 5 (cit. on p. 45).

*Walk score professional*. (2019). www.walkscore.com. (Cit. on pp. 89, 91, 98)

Walker, A. M., DeVito, M. A., Maestre, J. F., Siek, K. A., Kresnye, C., Jelen, B., Shih, P. C., Wolters, M., & Alqassim, M. (2019). Arc: Moving the method forward, In *Extended abstracts of the 2019 chi conference on human factors in computing systems*. (Cit. on p. 132).

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, In *Proceedings of the first workshop on nlp and computational social science*. (Cit. on p. 109).

Waycott, J., Vetere, F., Pedell, S., Kulik, L., Ozanne, E., Gruner, A., & Downs, J. (2013). Older adults as digital content producers, In *Proceedings of the sigchi conference on human factors in computing systems*. (Cit. on p. 82).

West, P., Giordano, R., Van Kleek, M., & Shadbolt, N. (2016). The quantified patient in the doctor's office: Challenges & opportunities, In *Proceedings of the 2016 chi conference on human factors in computing systems*. (Cit. on p. 110).

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis, In *Proceedings of hlt/emnlp 2005 interactive demonstrations*. (Cit. on p. 31).

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, ACM. https://doi.org/10.1145/3173574.3174230. (Cit. on pp. 15, 16, 49, 51, 126, 130)

Wright, J. D., & Devine, J. A. (1992). Counting the homeless: The census bureau's" s-night" in five us cities. *Evaluation review*, *16*(4), 355–364 (cit. on p. 14).

Wright, J. D., & Devine, J. A. (1995). Housing dynamics of the homeless: Implications for a count. *American Journal of Orthopsychiatry*, *65*(3), 320–329 (cit. on p. 14).

Xie, B., Watkins, I., Golbeck, J., & Huang, M. (2012). Understanding and changing older adults' perceptions and learning of social media. *Educational gerontology*, *38*(4), 282–296 (cit. on p. 82).

Yousef, O. (2018). Activists: Gang database disproportionately targets young men of color. Chicago Public Media Inc. https://www.wbez.org/stories/activists-chicago-gang-database-disproportionately-targets-young-men-of-color/693f6b8b-4252-48b8-9812-5d461c99e86a. (Cit. on p. 131)

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 194 (cit. on pp. 16, 126, 129, 135).

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089 (cit. on p. 21).

# Appendix

## 8.2. Demographic Survey

### 8.2.1. Please indicate your age.

50-59
60-69
70-79
80-89
90-99
100+

### 8.2.2. Please indicate your race or ethnicity.

White
Black or African American
American Indian or Alaska Native
Asian
Native Hawaiian or Pacific Islander
Middle Eastern
Other Ethnicity (write)

### 8.2.3. Are you Hispanic or Latino?

Yes
No

### 8.2.4. How would you describe the area where you grew up?

Urban
Suburban
Rural

### 8.2.5. How would you describe the area where you currently live?

Urban
Suburban
Rural

### 8.2.6. Please indicate your annual *household* income.

Less than $10,000
$10,000 - $14,999
$15,000 - $24,999
$25,000 - $34,999
$35,000 - $49,999
$50,000 - $74,999
$75,000 - $99,999
$100,000 - $149,999
$150,000 - $199,999
More than $200,000

#### 8.2.7. Please indicate the highest level of education you have completed.

Less than high school
High school graduate, GED, or equivalent
Some college or associate's degree
Bachelor's degree
Graduate or professional degree

#### 8.2.8. Which employment status best describes you?

Working full-time
Working part-time
On disability
Retired
Unemployment
Other (write)

#### 8.2.9. How would you describe your current living situation (check all that apply?)

I live alone
I live with a spouse or romantic partner
I live with family members
I live with roommates
I live in an assisted living facility
I live in a retirement community
I live in a nursing home
Other (write)

#### 8.2.10. How would you describe your political leaning?

Very liberal
Somewhat liberal
Moderate
Somewhat conservative
Very conservative

#### 8.2.11. Please indicate your gender.

Man
Woman
Nonbinary

## 8.3. Aging Experience Survey

*These questions were generated in an exploratory manner.*

**8.3.1. Do you consider yourself to be an older adult?**

Yes
No

**8.3.2. I have heard _____ about age discrimination.**

A great deal
Some
A little
Nothing

**8.3.3. In recent years I have been discriminated against or treated more negatively by others because of my older age.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.4. In recent years I have avoided certain social settings out of concern that I might be treated negatively because of my age.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.5. Age discrimination is too often excused.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.6. Age discrimination is a major problem in our society.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.7. I have had experiences with age discrimination that have caused me mental or emotional stress.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.8. Age discrimination is taken too seriously.**

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

**8.3.9. Have you or anyone you know ever seen or been the target of discrimination in the workplace based on older age?**

Yes
No

**8.3.10. Have you or anyone you know ever seen or been the target of discrimination based on older age while applying or interviewing for jobs?**

Yes
No

**8.3.11. Have you or anyone you know ever seen discriminatory depictions of older age in product marketing or advertisements?**

Yes
No

**8.3.12. Have you or anyone you know ever seen discriminatory depictions of older age in television shows or movies?**

Yes
No

*The next few questions discuss automated algorithms. An automated algorithm is a series of steps that a computer follows automatically to reach a set goal, such as calculating a credit score or showing specialized advertisements to specific online customers. Because the process is automated, it happens with little or no human control.*

**8.3.13. How acceptable do you believe it is for organizations to use age-related information in algorithms to show you targeted ads for products and services?**

Not at all acceptable
Not very acceptable
Somewhat acceptable
Very acceptable

**8.3.14. How acceptable do you believe it is for organizations to use age-related information in algorithms to review employment applications?**

Not at all acceptable
Not very acceptable
Somewhat acceptable
Very acceptable

**8.3.15. How acceptable do you believe it is for organizations to use age-related information in algorithms to review applications for social services?**

Not at all acceptable
Not very acceptable
Somewhat acceptable
Very acceptable

**8.3.16. Have you ever been aware of an automated algorithm used to review an application of yours for a job or employment?**

Yes
No

**8.3.17. Have you ever been aware of an automated algorithm used to present you with a product advertisement?**

Yes
No

**8.3.18. Have you ever been aware of an automated algorithm used to review an application of yours for a public or social service?**

Yes
No

## 8.4. Aging Anxiety Survey

(*Denotes reversed for scoring purposes)

### 8.4.1. Factor I: Fear of Old People

1. I enjoy being around old people.
3. I like to go visit my older relatives.
10. I enjoy talking with old people.
13. I feel very comfortable when I am around an old person.
19. I enjoy doing things for old people.

### 8.4.2. Factor II: Psychological Concerns

*5. I fear it will be very hard for me to find contentment in old age.
7. I will have plenty to occupy my time when I am old.
11. I expect to feel good about life when I am old.
16. I believe that I will still be able to do most things for myself when I am old.
18. I expect to feel good about myself when I am old.

### 8.4.3. Factor III: Physical Appearance

4. I have never lied about my age in order to appear younger.
9. It doesn't bother me at all to imagine myself as being old.
12. I have never dreaded the day I would look in the mirror and see gray hairs.
15. I have never dreaded looking old.
20. When I look in the mirror, it bothers me to see how my looks have changed with age.

### 8.4.4. Factor IV: Fear of Losses

*2. I fear that when I am old all my friends will be gone.
*6. The older I become, the more I worry about my health.
**8. I get nervous when I think about someone else making decisions for me when I am old.
*14. I worry that people will ignore me when I am old.
*17. I am afraid that there will be no meaning in life when I am old.

### 8.5. Age-Related Test Set

A lot of <older/younger> people looking for work will be suffering in the next few years.
Actually the fact that we all get <old/young> is a contributing factor, IMHO. It makes every joke and slight seem ironic! And therefore less serious or actionable.
Ageism is an interesting topic for me - I work in home care. I observe how often a doctor will dismiss symptoms with "you're just getting <old/young>".
Almost everyone <older/younger> than 55 needs glasses at least part of the time.
Also they need to lobby the Government to put policies in place that will help them and the following generations navigate <old/young> age more comfortabley
And all the while I'm thinking about the simplist of racism definitions – 'noticing a difference', and that aptly explains the separation of reactions that you so aptly described if the scoffer had been an '<old/young> fart'.
And I thought, in the United States, we put ankle bracelets on criminals and people under house arrest, not <old/young> people.
And it is shameful how <older/younger> people are treated by every aspect of American culture.
And just why is it that society (let alone the journalists) agrees that middle-aged and <older/younger> men can still be a "catch" but women can't?
And then, well, that day never arrived and one thing led to another until there was hardly any space left for an <old/young> woman to live in.
And yet two <old/young> people agreed of their own free will to star in the sketch.
Around here there aren't any special admission prices for <older/younger> people, but if there were it wouldn't bother me.
As Baby Boomers grow <older/younger>, there will be an increased need to meet their mobility needs, but Medicare won't be there for them.
As I grew <older/younger>, I began to fear solitude and being alone.
As much as I work on acceptance of getting <old/young>, I don't like it!
At the gym the other day, huffing and puffing away, I noticed a wispy-haired gent who I hoped is <older/younger> than even I am, whipping through a few weight routines.
Browsing for vibrators with a chirpy group of <older/younger> women offers some eye-opening reminders about aging, sex and camaraderie.
But almost every other number from the bureau makes it clear that while the economy may be improving, a substantial number of <older/younger> workers who lost jobs are still suffering.
But by teaching <older/younger> people how to use computers and the internet, we can go a long way toward helping to solve the problem.
But it is not <older/younger> folks who should be ashamed and embarrassed; it is the culture at large.
But, volunteer? Well, I have to turn down volunteer work. Everyone wants the "<old/young> lady" for that work! I find that annoying.
Cher is another example of really tragic work that lost the <old/young> lady she could have been.
Each of us is an amalgam of all of our previous 'selves', and I see this as a reminder that an <older/younger> person is not just an '<old/young> person', but an interesting individual with an individual history which has made them who they are today.
Eons ago in internet time, during the first year of this blog, there was a discussion over several posts about theories - reasonable and farfetched - on the well-known phenomenon of time appearing to speed up as we get <older/younger>.
Every one is telling the world, every day, what its really like to get <older/younger>.
Financial abuse is the illegal or improper use of an <older/younger> person's funds, property, or resources.
For goodness sake! Get over being <old/young>, relax and enjoy it.
For many people, however, these prove to be minor impediments and the practice of intimacy physical and emotional connectedness for many <older/younger> people continues to be pleasurable, rewarding, and fulfilling.
Four percent of Internet users 65 and <older/younger> say they use Instagram.
Getting <old/young> isn't easy and the little things seem to pile up
Growing <old/young> is an accomplishment.
Growing <old/young> is the definition of life and they are as handsome now - with their decades of living on display - as they were in their twenties.
Harry was a long-time family friend, too <old/young> by World War II to be drafted.
He and his corporate profits don't care about individuals - and many of them don't believe they will ever get <old/young>.
Here's one <old/young> white man who cheered. What wonderful progress our country has made in my lifetime.
His Edenization of long-term care environments has dramatically improved the quality of life for countless <older/younger> adults, engaging them in their own care creating active, home-like communities.

However, I think a lot of <old/young> people go with whatever the trend is without thinking much about it and as issues like gay marriage are more in the media, it becomes the "thing" for them to support the issue.
I agree that the ads for <old/young> people are ageist and only deal ailments. I often have thought don't <older/younger> adults buy clothes, electronics, food?
I always wondered why <older/younger> people would retire to area's they'd never lived in before.
I am 25 years <older/younger> than him.
I am always thrilled to find <older/younger> people blogging since they write well on a variety of topics and can also spell.
I am embarrassed by <older/younger> people who expect or feel entitled to be treated a certain way, simply because they managed to live to a certain age.
I am very afraid for my future as a <older/younger> single woman who is very much alone.
I call myself <old/young> because, at almost 64, what else fits?
I can imagine many things the <old/young> person might be thinking.
I don't mind being <old/young>.
I earned every line on this <old/young> face. I have no desire to be twenty or even thirty.
I enjoy talking with very <old/young> people.
I feel sorry for women who take that attitude about growing <older/younger>.
I guess lots of Daily Show viewers find <old/young> people maddening.
I have several friends who vehemently deny they are <old/young> or getting <old/young>.
I hope that subliminally tells them that the <oldest/youngest> generation is special in its way.
I know very few "<old/young>" people who don't personally know gay people.
I love seeing <older/younger>, non-professional women modeling clothes.
I never thought of myself as being <old/young> until I began (out of necessity) to take advantage of so-called entitlement programs.
I only discovered these artists for real when I was well on the road to becoming an <old/young> lady.
I suspect that the main treatment for <older/younger> depressed people is medication, which never results in attitudinal change.
I think I'm <old/young> enough to decide what I can eat!
I think the best thing about <old/young> age is learning not to give a damn what anyone thinks of me.
I think the reason for condescending attitudes towards the <old/young> is dread. Dread of getting <old/young> themselves and dying.
I think we resist looking <older/younger> because we don't want to be tossed on the scrap heap of our culture – our life skills, experience and contributions demeaned and devalued.
I try to deal with overt ageism with humor .....telling the person or persons who inflict it, they, too, will be <old/young> someday IF they are lucky.
I used to write a lot about fearing that I'd become one of those <old/young> women with too many cats.
I watched a woman get fired for being overweight as well as <older/younger>. Another one was fully supporting her husband who had lost his eyesight at work.
I wonder if this is what <old/young> feels like.
I, too, am confused about how <old/young> people are regarded. When did all of this seem to be okay to so many?
I'm just so much <older/younger> than him.
If <older/younger> people had written the script, it would have been much different.
If <older/younger> writers are more interesting, perhaps it has to do with accumulated years of experience.
In 1965, there were approximately 18 million people age 65 and <older/younger> in the United States.
In a system that does not value maintaining the whole, those of us who are <old/young> lose value quite early, long before we're actually a burden.
In the book department was a man who appeared to be quite <old/young>.
In these days, I'm glad I'm <old/young> and childless.
Is it just my imagination or do seniors in a group situation such as senior centers or senior living facilities act more like the stereotypical <old/young> people than do seniors who tend to stay away from those groups.
It also upsets me when I realize that society expects this from <old/young> people.
It gets harder to believe about ourselves as we grow <old/young>.
It is a mark of ingratitude to resent growing <old/young>.
It is a mistake to compare our lives to one another and anyway, I think <old/young> age is the great leveler.
It is a wonderful time in which to grow <old/young>.
It is all but impossible to find attractive womens clothing for bodies <older/younger> than about age 25.
It use to be that <older/younger> was wiser and better plus respected.
It's how, with every passing day, I look more like my mother in her <old/young> age and I'm not pleased about that.
It's insulting to the intelligence of most <older/younger> folk.

it's never easy to be the <oldest/youngest> person in a room.
It's starting to be a trend to lay off <older/younger> workers.
Just looking at the picture of Mr Shkreli, before I read the post about what he was doing I thought..well theres a smug looking <old/young> man with a lot of snark in his life.
Life and learning does not end in <old/young> age.
Like most 'isms', bias against <older/younger> folks has more to do with blindness and ignorance than bad intention or conscious cruelty.
Many <old/young> people feel that way.
Many <old/young> people who retired before computers were ubiquitous in the workplace have had to teach themselves the intricacies of these machines or rely on children and grandchildren.
Many <older/younger> workers relied on the growing value of their home as one kind of investment for their retirement.
Millions of <older/younger> adults means that change will occur,if everyone takes responsibility for his/her part.
Most <older/younger> folk, I speak to about this, are adamantly opposed to sharing their home with a roomate for company and to offset expenses.
Much has been discussed on this blog about what words are used to describe <old/young> people.
My <old/young> cheating dog of an ex-husband will just die from his foolishness when he finds out.
My Catholic faith dwindled as I grew <older/younger> and more disappointed that the Church did not take a stand on many human rights issues.
My mother said she hated reaching 80 because for the first time she felt <old/young>.
No one in West Virginia was afraid of growing <old/young> that I could see or sense.
Now however, you're a "70 something <old/young> codger barely able to tell reality from imagination. Feeling depressed yet Peter?
Now that I am <old/young> and don't give a shit how I look, I can pretty much do the same thing.
Now there is no question in my mind or in the minds of my acquaintances that I am <old/young>.
Now these people become <old/young> and are tagged as forgetful due to age as if that is the primary reason.
<Old/Young> age creeps up on you.
<Old/Young> age is worth waiting for.
<Old/Young> people's appearance contains so much lived life
<Old/Young> age can be rich, beautiful, and rewarding.
<Old/Young> age can prove to be a curse in lot many cases when with falling memory and declining health one is unable to take care of everyday situations.
<Old/Young> people's appearance contains so much lived life
<Older/Younger> adults must learn to be more pro-active rather than just taking it.
<Older/Younger> workers and all women were often excluded by laws limiting the amount that anyone could carry.
One day does blend into the next, and yes it is a privilege to grow <old/young>.
One day perhaps books with <older/younger> characters won't stand out in our minds because they are just normal.
One of the huge unspoken problems for <older/younger> people is finding transportation to and from the doctor's appointment–that becomes more of a stress than the appointment itself!
One of the most common complaints I hear from <old/young> people is about sleep or lack thereof.
One of the unavoidable things about getting <old/young> is that our public contemporaries the writers, actors, singers, songwriters, artists and other celebrities who help define our generation grow <old/young> with us.
Over the weekend, I came across a perfectly dreadful essay about how awful it is to look <old/young>.
Pat, I loved your comments! What a wonderful way to look at the "gift" of <old/young> age!!
People 65 and <older/younger> comprise about 13 percent of the U.S. population but account for 34 percent of all prescription medicine use and 30 percent of all over-the-counter (OTC) drug use.
People in the <older/younger> of these two groups are worse off than they were a year ago.
People joke about it because they know they have no choice; they will be <old/young> too someday. And I don't fault them for it.
People live to be far <older/younger> than that.
Perhaps that's cause being prejudice toward <older/younger> people comes so naturally in U.S. culture.
Perhaps this is something - among others - everyone must get through to eventually find a path to a fulfilling <old/young> age.
Sexism is nothing compared to growing <old/young>.
She should just be glad she is growing <older/younger>, since the alternative is not to our liking.
Since then, there have been continual reports of <older/younger> folks in long lines in the hot sun with nothing more useful than hope that the vaccine will not have run out when their turn comes.
Some (but not all) of the <older/younger> adults had a harder time overlooking unimportant information.
Some are lucky enough to have good health and find no issues in getting <older/younger>.

Still, I resist the limitations that an ageist society tries to place on me as an <older/younger> woman, and the fact that I must admit: some of them are actually based in fact.
That patronizing look of pity makes me want to jump over the counter, take them by their shirt fronts and inform them that if they're lucky they will be <old/young> someday too.
That's when I knew I was <old/young>.
The <older/younger> adults' brain scans showed activity in the same area.
The <older/younger> we get the fewer options we have to care for ourselves financially.
The <older/younger> worker loses their health and dental benefits, they lose out on contributing to their pension fund and social security (so their social security paycheck will be less at retirement).
The constant message that <old/young> people are expected to be slow and weak and forgetful is not a reason for the full-blown frailty syndrome.
The one and only thing I don't do because I am <older/younger> is buy a parrot or plant a tree.
The teacher was terrific, the price was right, the place close to home, the people in the class delightful. The problem? My shock on seeing my "<older/young>" classmates.
Then there are the less life-threatening afflictions of growing <old/young>, among them arthritis, sleep difficulties, hearing loss, vision problems, osteoporosis and others.
There are a lot of 'non-thinkers' out there and they get <old/young> too.
There are plenty of other people much <older/younger> than I.
There are thousands - maybe tens of thousands - of blogs written by <old/young> people.
They will fire you from your job or not hire you, because you are now <old/young>.
Think of it as a telephone party line (that's a joke for those of you <old/young> enough to remember party lines).
Think of the gazillion ads for wrinkle creams, age spot removers, health clubs, etc. that send out the message that only <old/young> is beautiful.
This <old/young> guy was 3 or 4 feet from the tide line and the tide was going out.
This economic collapse has been disastrous for too many <old/young> people.
This fearfulness should dissipate as boomer women, who worked in the big city and commuted and earned their own living, turn <old/young>.
This is a strange comment, above. Sounds like Gloria Steinem's being taken to task simply for getting <old/young>.
Those of us lucky enough to grow <older/younger> need to make the most of it.
To me, <old/young> age is always ten years <older/younger> than I am.
Until lately my attitude about getting <older/younger> was colored with worry about my body and brain becoming more and more disabled with time.
We know it anecdotally from readers we've heard from who've been blatantly discriminated against because they're <older/younger>.
We live in a culture that deliberately hides and ignores <older/younger> folks.
We males, esp. <older/younger> white males, get this all the time. We're used to it.
We need a similar push to get them into nursing homes, assisted living facilities, retirement communities and in homes of <older/younger> people who live independently on fixed incomes.
We of the <old/young> have had our share of the bad and sadness of life.
Well, I am <old/young> and I know plenty of gay people, and guess what, I like most of them about as well as I like anyone in the straight population.
When I started the book I thought all I needed to have a meaningful and healthy <old/young> age was to eat better and exercise.
When we get <old/young>, our ability to communicate sometimes is lessened and it does take an advocate to be there and be sure it's okay.
When we met, I asked him how <old/young> he was, but he wouldn't tell me. He said, Then you will start treating me like an <old/young> person.
When you're 50 going on living, it's called <old/young> age.
Whenever I think ageism is pointed in my direction, I have to honestly assess how I thought of <older/younger> people when I was 20 or 30
While both men and women do hit that barrier when suddenly they're "too <old/young>" to be hired, or paid attention to, or be respected as effective agents... women hit it harder, and earlier.
Why is it, will someone please explain, that when government budgets are being hammered out, the first place legislators go for cuts is <old/young> people, the ones who have paid into and earned their pensions and Social Security?
You are so right especially about how egregious it is that many <older/younger> people can't get health insurance at any price.
You never know when it's going to happen to you and no one can put a date on it for you because we grow <old/young> in appearance (and other ways too) at different rates.

You're too <old/young> to get a teaching job.
Your description of how attendees ignored an <old/young> woman who was not like them has EVERYTHING to do with wrinkle creams.
Your post reminded me of a video on TV of a robber caught by a video camera in a New York hallway, attacking an <old/young> lady.