Measuring the Immeasurable

Perspectives on Best Practices When Operationalizing Machine Learning Fairness for Recommender Systems

written by

Jessie J. Smith

B.S., California Polytechnic State University, 2018

M.S., University of Colorado Boulder, 2023

A dissertation submitted to the Faculty of the Graduate School of the University of Colorado Boulder in partial fulfillment of the requirement of the degree of Doctor of Philosophy

Department of Information Science, 2025

Committee Members:

Dr. Casey Fiesler (co-chair)

Dr. Robin Burke (co-chair)

Dr. Jed Brubaker

Dr. Jennifer Wortman Vaughan

Dr. Michael Madaio

Abstract

Jessie J. Smith (PhD, Information Science)

Measuring the Immeasurable: Perspectives on Best Practices When Operationalizing Machine Learning Fairness for Recommender Systems

Dissertation advised by Dr. Casey Fiesler and Dr. Robin Burke

Incorporating fairness into the design of machine learning (ML) systems is a central component of responsible technology design. However, in machine learning, and particularly within the subdomain of recommendation systems, applying concepts of fairness requires attention to various stakeholders' complex and often-conflicting needs. Since fairness is a socially constructed, theoretical, and context-dependent construct, there are numerous definitions that span multiple disciplines. Still, it is rare for machine learning researchers to develop their metrics in close consideration of their social context, or with the input of those who are most impacted by these systems. More often, standard definitions are adopted and assumed to be applicable across contexts and stakeholders. This complexity of fairness definitions and contexts coupled with the proliferation of fairness metrics in research literature have led to a challenging decision-making space for practitioners to navigate. In addition, the practical incentives, motivations, and constraints of doing fairness work in industry settings limit what is possible for practitioners.

In this dissertation, I confront these challenges by uncovering potential best practices and

methods for operationalizing fairness in multistakeholder recommender systems. Drawing on five qualitative research studies involving over 100 end-users, item providers, practitioners, and ML fairness experts, I examine the intricate process of translating subjective perspectives and theoretical fairness frameworks from moral philosophy into empirical investigations suitable for observable settings. I define "pragmatic fairness" as the path from business constraints to imperfect (though satisfactory) fairness evaluations, and I explore how fairness evaluation can still be successful within this practical setting. Finally, I discuss the tradeoffs between best practices and realistic practices when operationalizing fairness in industry, and I point towards promising future work to help bridge this gap.

Acknowledgments

Seven years ago I did not know what a PhD was, and now–with the help, guidance, support, and care of so many incredible people–I have made it to the end of this journey. Here is where I share my gratitude for these people.

Thank you Alex Boyd for teaching me what a PhD is, and encouraging me to go for it in my final year of undergrad. Thank you Hanna Wallach and Abbie Jacobs for answering my cold-call emails and taking the time to introduce me to the field of Information Science all those years ago when I was bright-eyed and bushy-tailed and had no idea what I was doing applying to PhD programs. Thank you Dennis Sun for introducing me to the concept of machine learning fairness and starting me on my path in this career; thank you for taking a chance on me and being my first research advisor. Thank you Nicholas Sakellariou–your computer ethics class changed my life in so many ways, and made me realize I wasn't done wanting to learn yet.

Thank you Casey Fiesler–not only for your immense support over the years of this degree and as my PhD co-advisor, but also for writing your medium articles and doing public scholarship on tech ethics, if not for that, I would have never found the Information Science program at CU. Thank you Robin Burke, my other co-advisor, for all of your guidance and countless hours of support as well; thank you for introducing me to the field of recommender systems. Thank you to both Casey and Robin for teaching me

practically everything I did not know about Academia coming into this degree–I could not imagine finishing this degree without both of your mentorship, guidance, and replies to my incessant Slack messages. Thank you also to the rest of my dissertation committee: Jed Brubaker, Jenn Wortman Vaughan, and Michael Madaio. Jed, I will take many lessons I learned in your theory class with me through the rest of my career. Jenn, thank you for seeing my potential so early on, and mentoring me through my first PhD internship and beyond. Michael, thank you for the immense support you've given me in this final push of my PhD–it has been an absolute joy to work alongside you.

Thank you also to the many other Academic/Industry collaborators and general supporters throughout this journey. To name a few: Samantha Dalal, Nasim Sonboli, Julia Dean, Josh Meyers-Dean, Wesley Deng, Joshua Paup, Solon Barocas, Lex Beattie, Nicole DeCario, Jesse Dodge, Michael Ekstrand, Chris Dulhanty, Morgan Scheuerman, Emily Bender, Henriette Cramer, Alex Beutel, Kate Sousa, and Zoya Bylinski–your research, collaborations, and emotional support have been incredibly invaluable to me.

Thank you to the Internet Rules Lab and That Recommender Systems Lab for all the discussions we've had on tech ethics over the years, and all of the invaluable feedback you've given me on my work.

Thank you to everyone who participated in my research studies. Your stories are what made this work possible.

Thank you Shamika Klassen for doing SciFi in Real Life with me and being such an inspirational PhD cohort friend, and thank you Dylan Doyle for doing The Radical AI Podcast with me–my career is forever changed for the better because of that project.

Thank you Claudia Macedo for being my rock through it all. And thank you to Yeng Tan, my biggest dissenter (I promised him I would put that in here). Thank you Anas Buhayh for consistently inspiring me every day through work, hobbies, and friendship;

I love you.

I of course cannot write an acknowledgements section without thanking my wonderful family. Thank you Mom for all of your endless emotional support and love, and for being my biggest career coach and cheerleader. Dad, thank you for convincing me to take my first data science class 8 years ago even though I was scared to, and thank you for sending me that article on AI's impact on society all those years ago, I didn't realize I was looking into my future at the time. Thank you to Eric and Wesley for the laughs, the phone calls, and being the best siblings I could ask for.

And finally, thank you to all of my friends in Boulder, Colorado for fulfilling my life outside of work; being surrounded by a supportive community has helped me philosophize, create, and most importantly–to write. It takes a village, and I am immensely grateful for the one I've landed in.

Thank you, thank you, thank you–everyone named and unnamed here, for the support over the last (almost) decade it has taken to get here. I hope you enjoy this dissertation, this is only the beginning.

Contents

1 Introduction **1**

2 Background & Literature Review **8**

2.1 A Brief History of Fairness in Machine Learning 8

2.2 Theoretical Foundations and Definitions 11

2.2.1 What is "Fairness"? 11

2.2.2 Operationalizing Fairness in Recommendation 16

2.2.3 Measurement Theory 21

2.2.4 Designing for Human Values 28

2.2.5 Ethical Theories 30

2.3 Applying ML Fairness in Industry Settings 34

3 Methods **40**

3.1 Study #1: The End-Users' Perspective 40

3.1.1 Kiva as a Case Study 42

3.2 Study #2: The Providers' Perspective 43

3.2.1 Participants 43

3.2.2 Focus Group Design 44

3.2.3 Data Analysis 47

3.3 Study #3: The Practitioners' Perspective (Kiva) 48

3.3.1 Participants 48

3.3.2 Data Collection 49

3.3.3 Data Analysis 49

3.4 Study #4: The Practitioners' Perspective (Spotify) 50

3.4.1 Designing the Prototype 51

3.4.2 Interviews 54

3.5 Study #5: The Experts' Perspective 55

3.5.1 Participants 56

3.5.2 Data Analysis 59

3.6 Methodological Limitations 59

3.6.1 Positionality Statement 62

4 The Users' Perspective **64**

4.1 Motivation 64

4.2 Study #1: The End-Users' Perspective 68

4.2.1 Introduction 68

4.2.2 Results . . . 69
4.2.3 Discussion . . . 77
4.3 Study #2: The Providers' Perspective . . . 79
4.3.1 Introduction . . . 79
4.3.2 Results . . . 80
4.3.3 Discussion . . . 94
4.4 Conclusion . . . 97
5 The Practitioners' Perspective 100
5.1 Motivation . . . 101
5.2 Study #3: Fairness at Kiva . . . 105
5.2.1 Introduction . . . 105
5.2.2 Results . . . 106
5.2.3 Discussion . . . 118
5.3 Study #4: Fairness at Spotify . . . 124
5.3.1 Introduction . . . 124
5.3.2 Results . . . 125
5.3.3 Discussion . . . 132
5.4 Conclusion . . . 138
6 The Experts' Perspective 141
6.1 Motivation . . . 141
6.2 Study #5: The ML Fairness Experts' Perspective . . . 146
6.2.1 Introduction . . . 146
6.2.2 What is Fairness and What is the Goal of Evaluation? . . . 147
6.2.3 The Path From Constraints to Satisficing . . . 158
6.2.4 Pragmatic Fairness in Action . . . 164
6.2.5 Improving Pragmatic Fairness . . . 170
6.3 Conclusion & Pathways Forward . . . 176
7 Bigger Picture 178
7.1 Takeaways & Implications . . . 179
7.2 Where Do We Go From Here? . . . 188
7.3 Summary . . . 194
7.4 Concluding Remarks . . . 195
Bibliography 198
Appendix 221

List of Tables

1.1 The Five Studies of This Dissertation 7

2.1 Evaluating Construct Validity 25

3.1 Study #2 Participants (Content Creators) 45
3.2 Study #2 Participants (Dating App Users) 46
3.3 Study #4 Participants 55
3.4 Study #5 Participants 58

4.1 Study #2: Participants' Fairness Goals, Definitions, and Metrics 85

5.1 Study #4: Consumer-Side Fairness Metric Categories 135
5.2 Study #4: Provider-Side Fairness Metric Categories (Part 1) 136
5.3 Study #4: Provider-Side Fairness Metric Categories (Part 2) 137

6.1 Study #5: Examples of Pragmatic Fairness 165

List of Figures

3.1 Study #4 Interview Process 54

4.1 Study #1 Explanation Design 1 74
4.2 Study #1 Explanation Design 2 75
4.3 Study #1 Explanation Design 3 75

5.1 Study #3: Fairness Logics, Tensions, and Stakeholders 114

6.1 Decision-Making Stages for ML Fairness Evaluation 142
6.2 The Futures Cone 145

Chapter 1

Introduction

> **"Measures are more than a creation of society, they *create* society"**
> (Alder, 2003).

In the dynamic field of machine learning (ML), the intersection of the empirical and the theoretical is a pivotal research frontier. Over the past few decades, computer scientists have revolutionized the way we build, measure, and optimize ML models to achieve specific objectives. Through empirical studies, ML researchers have identified numerous ways to operationalize theoretical concepts such as "accuracy" and "error." In the last decade, one challenging objective has surfaced in the field: *fairness*.

The concept of fairness has been a topic of debate for millennia. Defining what it means to treat people fairly is a daunting task that lacks a single, clear-cut solution. Selbst et al. (2019) said that *"fairness and discrimination are complex concepts that philosophers, sociologists, and lawyers have long debated. They are at times procedural, contextual, and politically contestable, and each of those properties is a core part of the concepts themselves."* Operationalizing a theoretical construct such as fairness poses an inherently complex challenge. It requires researchers to empirically observe something that is funda-

mentally unobservable; to measure the immeasurable. Additionally, fairness is a concept that remains contested and context-dependent. To attempt to measure and observe such a construct requires making many assumptions and reductions that can never holistically capture the construct itself.

In the realm of algorithms—technical systems that comprise of inputs, instructions, and outputs (Hogan, 2015)—defining what it means for these systems to treat people fairly is just as challenging. A key aspect of this complexity is that while there is no unified definition of fairness, nor any consistent method to achieve fairness, algorithmic systems must still make decisions about which types of fairness to measure and optimize for. As the adage goes, ***you can't change what you can't measure.*** This intersection between the theoretical and the empirical represents a pivotal juncture where math meets morality, and is a focal point of this dissertation.

Developing robust scientific methodologies to translate subjective, independent experiences into scientifically observable, measurable, and optimizable constructs is an essential part of incorporating fairness into machine learning systems. However, it is one aspect of this research discipline that has received little attention. What is more common is for researchers to assume that an accurate definition of fairness has been chosen (accurate meaning that it appropriately matches the context of the system and its stakeholders), and to create metrics and algorithms that correspond with these definitions. However, since fairness as a concept is contextual, contested, and stems from conflicting theoretical understandings (Kleinberg et al., 2016), optimizing algorithms for one definition of fairness may benefit some and harm others.

One fascinating application of ML that has begun to explore incorporating fairness objectives is the domain of recommender systems. These often personalized systems leverage algorithms to recommend content, items, or information that matches users' perceived preferences. However, previous work has highlighted how personalized systems might also

lead to unintentional harm, such as degenerate feedback loops (Jiang et al., 2019; Ribeiro et al., 2020), sexist stereotyping (HRW, 2018), or racial bias (Angwin and Parris, 2016; Asplund et al., 2020). This realization has resulted in calls to action for industry to audit, understand, and mitigate the potential "unfairness" of their recommendation systems. However, guidance that informs practitioners about how to operationalize fairness in their systems is largely lacking.

Many fairness definitions have been introduced and explored by the recommender system research community (e.g., (Ekstrand et al., 2022; Deldjoo et al., 2022; Kuhlman et al., 2021; Patro et al., 2022; Lazovich et al., 2022)). **Fairness definitions** (or objectives, goals) consist of different restrictions, assumptions, and requirements that must be met in order for a machine learning model to be classified as "fair" under that definition. Note however that a model itself being classified as fair does not guarantee fairness of its wider context. This combination of requirements for a given fairness definition can be referred to as **fairness constraints**. Each fairness definition has associated constraints that might differ depending on the context of the system and its users, the goals of stakeholders, and aspects such as the structure of training or evaluation data. Thus, defining and refining what constraints need to be met for a given fairness context can be a complex process that involves experts from different disciplines (Madaio et al., 2020, 2022). For example, a system that recommends jobs to job-seekers may have very different fairness considerations than one recommending books to readers. Both systems might implement a similar metric, but their context, constraints, and goals might differ substantially.

Application of these nuances has been limited, in part because of an overly-simplistic formulation of fairness (Keyes et al., 2019) and, in part, because incorporating ML fairness research from academia into industry contexts is not straightforward. For example, with few exceptions (Kearns et al., 2019, 2018; Hashimoto et al., 2018), published research in ML often considers only one protected group of stakeholders around which fairness should

be considered, an unrealistic constraint for most applications.

As a consequence, results from fairness-aware machine learning research often lack relevance for ML practitioners (Cramer et al., 2019; Rakova et al., 2021). This is particularly true of systems that apply fairness to personalized recommendations, even though these systems are critical to how millions of individuals access news, shopping, social connections, and employment opportunities. It is important to situate fairness interventions in these specific application contexts.

Alongside the lack of guidance for practitioners, another important gap has emerged in fair ML scholarship: there is very little research that designs fairness interventions *with* the real users and stakeholders of the systems who will be impacted by fairness operationalizations. One of the core tenets of user-centered, human-centered, and community-centered research is to incorporate impacted users into the design process of technologies. In the case of fairness, it is rare that users are involved in the process of scoping fairness goals, and developing protocols to measure and mitigate unfairness in ML systems.

My dissertation addresses this disconnect between the operationalization of fairness in machine learning and the users it directly impacts. Additionally, it delves into the practical challenges practitioners encounter when implementing fairness initiatives in real-world recommendation contexts. While I deeply admire the aspirational visions of fairness—how it can be measured, evaluated, and operationalized to altruistically improve societal outcomes—I also acknowledge that fairness is inherently subjective and context-dependent, rendering it impossible to fully "achieve" or "ensure." As Fazelpour and Lipton (2020) argue, many fairness metrics are rooted in ideal theory, which assumes a perfectly just world and often neglects the nuanced challenges of practical application. Instead, fairness work often requires what Simon (1978) term *satisficing*—seeking satisfactory, rather than perfect, outcomes within the constraints of limited time, information, and cognitive resources. Practitioners must grapple with measuring an immeasurable concept and ad-

dressing a fundamentally unsolvable problem, all while making decisions and implementing interventions. Drawing inspiration from Amartya Sen's distinction between *nyaya* (ideal justice) and *niti* (incremental justice) (Sen, 2009), I situate my work at the intersection of idealistic and pragmatic approaches, advocating for actionable, incremental progress as a pathway toward fairness.

Ultimately, I believe fairness work requires striking a balance between working within existing systems and envisioning how those systems might someday be transformed. This dissertation navigates that balance by bridging idealistic aspirations and pragmatic realities, offering a path forward that integrates both visionary and practical approaches to ML fairness.

My overarching research questions for this dissertation are:

- **RQ1:** How do users think that fairness should be operationalized in real-world recommender systems?
- **RQ2:** What do practitioners know and need when operationalizing fairness in real-world recommender systems?
- **RQ3:** What are potential best practices for operationalizing fairness in real-world recommender systems, keeping in mind the practical constraints of this setting?

For this dissertation, I completed 5 studies total, two for RQ1, two for RQ2, and one for RQ3, as shown in Table 1.1. For reference, each of my participant populations in my dissertation can be defined as follows:

- **Users** are anyone who interacts with a recommender system and/or who might be impacted by the system. This includes both consumers and providers. **Consumers** or **End-users** are people who are recommended items (e.g., Spotify listeners, Netflix

viewers, lenders on Kiva.) **Providers** are people who provide items to be recommended (e.g., content creators on YouTube or TikTok, musicians who post their music on Spotify, borrowers on Kiva).

- **Practitioners** are industry professionals who work with and around machine learning technologies. This includes but is not limited to: machine learning engineers, data scientists, project managers, program managers, PR managers, business analysts, and other employees of a company that influence machine learning system design.
- **Experts** are people who have years of experience working with machine learning fairness topics. This can include practitioners, and can also include academics who have published peer-reviewed papers on machine learning fairness.

To address RQ1, I first conducted a motivating, formative study that explored what kinds of fairness interventions users wish practitioners would incorporate into their recommender systems, and how they would like to be informed about these decisions (Sonboli, Nasim and Smith, Jessie J. et al., 2021). I then conducted focus groups with recommendation providers from the domains of dating and content recommendation, to: (1) learn about their lived experiences of unfairness in recommendation/ranking; (2) map their lived experiences into fairness goals, definitions, and metrics; and (3) to assess the ability to co-design fairness metrics with impacted stakeholders (Smith et al., 2024).

To address RQ2, I conducted two research studies that explore the perspectives of practitioners. One study explored the challenges that practitioners face when defining fairness as it relates to their organization's values and goals (Smith et al., 2023c). The other study explored the challenges that practitioners face when determining which metrics to use to empirically measure a chosen definition of fairness in their recommender systems (Smith et al., 2023a).

To address RQ3, I conducted an interview study with ML fairness experts to better un-

Table 1.1: The five studies included in this dissertation, and hyperlinks to their associated Chapters and Sections in this document.

RQ#	Theme	Study #	Study Name	Citations
RQ1	The Users' Perspective [4]	Study #1	Fairness and Transparency in Recommendation: The Users' Perspective [4.2]	Sonboli, Nasim and Smith, Jessie J. et al. (2021), Smith et al. (2020)
		Study #2	Recommend Me? Designing Fairness Metrics with Providers [4.3]	Smith et al. (2024)
RQ2	The Practitioners' Perspective [5]	Study #3	The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation [5.2]	Smith et al. (2023c)
		Study #4	Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective [5.3]	Smith et al. (2023a), Smith and Beattie (2022)
RQ3	The Experts' Perspective [6]	Study #5	Pragmatic Fairness: Evaluating ML Fairness Within the Constraints of Industry [6.2]	*Submitted for review at the FAccT 2025 Conference*

derstand the practical constraints of conducting fairness evaluations in real-world settings, and to explore best practices within the bounds of these constraints.

My goals with this dissertation work are the following: (1) to better understand users' lived experiences of unfairness in recommendation, and how they would like fairness evaluations to capture their experiences and needs; (2) to understand the practical constraints of operationalizing ML fairness in the real world, including the challenges and needs of practitioners; and (3) to explore what best practices could be for real-world fairness evaluation, what the limitations of this work are, and how we can improve these processes in the future.

Chapter 2

Background & Literature Review

In this section, I describe the current state of fairness in the discipline of machine learning, including the historical context that it is situated within. I explore some of the current methods and promising research directions for operationalizing fairness, especially with respect to the domain of recommendations. Finally, I describe how this previous work has and must act as a bridge between academic research and industry contexts, and how this background informs and motivates my dissertation.

2.1 A Brief History of Fairness in Machine Learning

In the late 1960's research on fairness began to explore beyond theoretical debates and extended towards formal definitions of fairness. This involved defining which population subgroups to measure fairness for, and grappling with the ideas that certain fairness definitions are incompatible with one another, or that certain quantitative fairness definitions are too limited (Hutchinson and Mitchell, 2019). Although more robust methods for measuring fairness (e.g., development of mathematical and statistical criteria) were intro-

duced in the 1970's, many of these efforts were abandoned after *"different and sometimes competing notions of fairness left little room for clarity on when one notion of fairness may be preferable to another"* (Hutchinson and Mitchell, 2019). It wasn't until 2008 when Pedreschi et al. published *Discrimination-aware Data Mining* that the language of "discrimination" and "fairness" were reintroduced into the computer science discipline (Pedreshi et al., 2008).

Immediately following, in the 2010's, fairness research found a new resurgence, this time within the discipline of machine learning, most notably after the publication of the oft-cited paper, *Fairness Through Awareness* (Dwork et al., 2012). In this publication, the authors introduced a framework for fair classification that included a task-specific fairness metric and an algorithm to optimize a classifier for this metric. This seminal work acted as a foundation for fair machine learning research for multiple reasons, including: (1) it began the still occurring debate between notions of individual versus group fairness (Binns, 2020); and (2) it was the first attempt to quantify and optimize a machine learning model for fairness as a potential model objective.

Around this same time, another seminal piece was published: *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon* (Boyd and Crawford, 2012). In this work, the authors described how the era of Big Data brings opportunities for us to assess the assumptions, values, and biases that might accompany data-driven technologies such as machine learning models. Several years later, Barocas and Selbst published *Big Data's Disparate Impact*, which additionally confronted the 'non-neutrality' of data-driven algorithms, particularly as they might interact with American antidiscrimination law (Barocas and Selbst, 2016). One year later, Cathy O'Neil's book, *Weapons of Math Destruction*, similarly brought to light how machine learning technologies can easily reinforce pre-existing societal inequalities (O'Neil, 2017).

All of these seminal works had something in common—they all at one point or another

noted that social phenomena cannot always be quantified. *"There are a number of practical limits to what can be accomplished computationally"* (Barocas and Selbst, 2016).

Despite these warnings, in the mid-2010's, a slew of scandals began to gain media attention related to various ways that algorithms were treating certain groups of people unfairly. Most notably, and likely most often cited, is the case of COMPAS. In this scandal, ProPublica brought to light that Northpointe's recidivism risk prediction algorithm was unfairly discriminating against black defendants (Angwin et al., 2016).

Reasonably, academics and practitioners alike began to explore potential solutions to these kinds of problems. During this time, machine learning researchers and practitioners found new motivation to quantify and optimize for the value of fairness in their ML systems (e.g., (Hardt et al., 2016; Chouldechova and Roth, 2020; Ustun et al., 2019)). This resulted in a proliferation of fairness metrics and associated definitions that could be used to "ensure" that fairness had been "achieved" for a given model and context (most often classification models and most often in a credit scoring context) (Narayanan, 2018), and eventually an entire book was written which was dedicated to operationalizing fairness in machine learning (Barocas et al., 2018).

While this research was groundbreaking and motivated the beginning of the ML fairness research discipline, several major gaps still exist; when it comes to practically doing ML fairness work, practitioners are still largely unaware of how to scope their fairness goals, define fairness for their context, and empirically measure these specific notions of fairness in the first place. In short, even when fairness is a priority, practitioners aren't sure where to begin.

2.2 Theoretical Foundations and Definitions

2.2.1 What is "Fairness"?

Fairness is a complex, contested, contextual, and theoretical construct (Jacobs and Wallach, 2021). Fairness has been defined in the dictionary as *"a lack of favoritism toward one side or another"* (Merriam-Webster, 2002). Policies, such as the Civil Rights Act of 1964, have defined fairness as *"not discriminating on the basis of race, color, religion, sex or national origin"* (Civil Rights Act, 1964). Machine learning literature has similarly defined fairness as *"the absence of discrimination"* (Castillo, 2019). It is important to note that discrimination in its own right is not inherently value-laden; "discrimination" as a concept means to differentiate between two things, and is only "wrongful" based on the cultural context in which it is occurring (Selbst et al., 2019; Hellman, 2008). Discrimination as it relates to fairness is the more insidious form of differentiation, when it becomes *"the act of making unjustified, prejudiced distinctions between people based on the groups, classes, or other categories to which they belong or are perceived to belong"* (Wikipedia, 2023).

However, fairness is not only defined through the lens of discrimination, it is also often defined through the lens of bias, such as Schiff et al. (2021) who defined fairness in machine learning models as *"minimizing bias."* Bias itself is also a challenging word to define, with societal bias and statistical bias often getting conflated. Societal bias is not something that can be 'solved' or entirely removed from datasets or from machine learning systems, however statistical bias can be measured, evaluated, and optimized for. Studies have stated that alleviating the effect of statistical bias can be one method to build more fair ML models (Kamishima et al., 2011; Fish et al., 2016; Kamiran and Calders, 2009). However, although bias may be a source of unfairness, there are other factors that can contribute to unfairness in machine learning—such as discrimination, unequal error rates,

stereotyping, and algorithmic harm—which may be related to, but separate from bias (Mehrabi et al., 2021; Chouldechova and Roth, 2020). Thus, using the terms 'bias' and 'unfairness' interchangeably may not always be appropriate.

The discipline of philosophy has also generated its own definitions of fairness, including fairness as forms of equality, equity, egality, liberation, and justice (Gooden, 2015; Rawls, 2001). Rawls (2020) recently described fairness as a form of justice by asking, *"what liberties are necessary to protect individuals in the development and pursuit of the conception of the good life, given the specific sociotechnical character of the society in which they live?"*

Rawls (2001) previously described that justice-based fairness shall only permit inequalities if they are to the advantage of the least well off. These ideas are also in line with egalitarianism, the concept that all people are equal and deserve equal outcomes. Inspired by egalitarianism and Rawls' fairness as a form of justice is the notion of equality of opportunity—which posits that everyone must exist on a *"level playing field"* when receiving a benefit of some kind (Arneson, 2002).

Fairness itself is also a form of **political liberalism** in that it fits into various and conflicting comprehensive doctrines with overlapping consensus between them (Rawls, 1991). Despite these overlaps, attempts have been made to categorize fairness into various types, particularly in the domain of machine learning.

One categorization of fairness was recently made by responsible tech scholar Meredith Broussard, where she described the difference between "social fairness" and "mathematical fairness."

> "Social fairness and mathematical fairness are different. Computers can only calculate mathematical fairness. This difference explains why we have so many

> problems when we try to use computers to judge and mediate social decisions. Mathematical truth and social truth are fundamentally different systems of logic. Ultimately, it's impossible to use a computer to solve every social problem. So, why do so many people believe that using more technology, more computing power, will lead us to a better world?" (Broussard, 2023).

Fairness, however, is deeply culturally situated and cannot be universally defined without acknowledging the profound influence of context. What one person or community may consider fair could be entirely different from what another deems just, influenced by factors such as cultural upbringing, societal norms, geographical location, historical context, and even the specifics of a situation. Ethical frameworks such as Ubuntu, an African philosophy emphasizing interconnectedness and communal well-being, underscore this variability. Ubuntu ethics focuses on the idea that a person is a person through other persons, prioritizing relationships and collective harmony over individualistic notions of fairness. This perspective challenges Western rationalist fairness paradigms, which often emphasize abstract rules or principles, by demonstrating that fairness can instead emerge from the dynamics of relationships, situational nuances, and communal values (Emmanouilidou, 2018; Amugongo et al., 2023; Farao et al., 2024). I explore these ideas further in Section 2.2.5.

Recognizing fairness as inherently contextual reminds us that achieving fairness in any system, especially in global applications like machine learning, requires sensitivity to cultural and situational diversity. It is clear that fairness has no single definition, nor is it confined by the borders of a single discipline. The question, *what does it mean to treat someone fairly?* is one that has many different answers, all of which can be equally "correct" or "incorrect" depending on the cultural context and the subjective definition of fairness for that given context.

In the domain of machine learning (ML), fairness is often related to the inputs and outputs of ML technologies. One example related to inputs is that training datasets are often ridiculed when they lack a fair representation of the users who will be impacted by the ML model they are trained for. One example related to outputs is that machine learning algorithms are often the arbitrators of important outcomes for real people, such as predicting someone's likelihood of recommitting a crime, predicting someone's credit worthiness, or selecting whether someone's job application is good enough to warrant an interview. Incorporating fairness into machine learning outputs is thus equally important.

In the domain of recommendation, fairness is also an important consideration. Recommender systems are used as a way to filter information to people. Fairness in recommendation then, considers questions such as: Are we fairly representing the information space? Are we giving all item providers an opportunity to be seen or engaged with? And are we recommending items to users in a way that fairly distributes information or resources to those users? (Ekstrand et al., 2022).

All of these "fairness problems" face a similar challenge: there is no single, perfect solution that can address them. Each of these problems might require prioritizing a certain group of stakeholders, defining fairness in a different way, and measuring and optimizing for fairness in a different way. There also might be a dozen equally appropriate approaches to defining and measuring fairness for each of these contexts.

Chouldechova and Roth (2020) describe that *"most of the literature on fair [machine learning] focuses on statistical definitions of fairness. This family of definitions fixes a small number of protected demographic groups G (such as racial groups), and then ask for (approximate) parity of some statistical measure across all of these groups."* This approach is also related to **disaggregated evaluation**, where the goal is to uncover performance disparities by evaluating and comparing ML system performance between demographic groups. This framing of 'statistical definitions of fairness,' is also in line with Broussard's

category of 'mathematical fairness,' (Broussard, 2023). The increased hype to research fairness in machine learning over the last decade has been largely in line with these statistical/mathematical notions of fairness—where the goal is to empirically measure the fairness or unfairness of a machine learning system. This hype has led to a proliferation of fairness definitions and metrics that machine learning practitioners can use to 'audit' their algorithmic systems (Barocas et al., 2018; Narayanan, 2018).

However, the complexity of fairness definitions and the proliferation of fairness metrics in research literature have led to a complex decision-making space. This environment has made it challenging for practitioners to operationalize and pick metrics that work within their unique context (Verma and Rubin, 2018; Smith et al., 2023a; Richardson et al., 2021; Deng et al., 2022). These challenges have also contributed to a widening gap between the research community and practitioners concerning the availability of fairness metrics versus the ability to put them into practice.

How, then, do practitioners realistically make these kinds of decisions when incorporating fairness into their machine learning systems? Numerous previous studies have highlighted that practitioners need more guidance when it comes to doing practical ML fairness work (Beattie et al., 2022; Lee and Singh, 2021; Deng et al., 2022; Richardson et al., 2021; Holstein et al., 2019; Madaio et al., 2022, 2020; Wang et al., 2023; Smith et al., 2023a). In a domain where there are no "correct" answers (but decisions must still be made), how can practitioners make decisions in a way that aligns with best practices and keeps users at the center of their design decisions?

It is important for me to note that this dissertation work does not seek to define fairness in the traditional sense; there will be no single framework or protocol that is purported to be "the solution" or "the definition" of fairness. In contrast, this dissertation work seeks to explore all of the potentially infinite definitions of fairness that apply to recommender system fairness work; keeping in mind the complex, interdisciplinary background

of this concept, and with special focus on helping practitioners come up with their own operationalizations as they relate to their specific use-cases.

2.2.2 Operationalizing Fairness in Recommendation

> "Whether we consider a data science project fair often has as much to do with the formulation of the problem as any property of the resulting model," (Passi and Barocas, 2019).

In this work, I define **qualitative fairness objectives** as fairness goals, such as those surfacing from an algorithmic impact assessment of a machine learning system (Madaio et al., 2020; Metcalf et al., 2021). In contrast, **quantitative fairness objectives** reflect quantitative definitions of fairness found in the literature, which often are accompanied by objective-specific fairness metrics (such as demographic parity or top-k fairness) (Verma et al., 2020).

A common starting point for measuring fairness in a machine learning system is to declare what type of unfair impact or *harm* to measure. Crawford (2017) previously categorized harms in machine learning as two broad types: (1) harms of **allocation**; and (2) harms of **representation**. In recommender systems, harms of allocation might refer to a system's unfair distribution of attention or exposure of items, while harms of representation might refer to a system's unfair representation of reality for a given information space (Ekstrand et al., 2022).

It is important to note that allocative and representative harms can occur for both consumers and providers of recommender systems. For example, consider the scenario of a job recommender. Allocative harm could impact those who receive recommendations if females are only recommended lower paying jobs than males; allocative harm could also

impact those who provide recommended items if certain employers' job ads are not being delivered to appropriate candidates at the same rate as another similar employer.

Harm, impact, and unfair treatment are all related in machine learning systems and, in some cases, can be empirically observed through proxies. Thus, after determining a potential harm to target, practitioners will need to "scope" an appropriate fairness proxy to measure for their context. These proxies are necessarily reductionistic, since fairness is a subjective and non-observable construct (Jacobs and Wallach, 2021). Nevertheless, selecting "good" proxies is a critical step in fairness operationalization, since they serve as the most effective stand-in for the concept of fairness itself. As Passi and Barocas (2019) describe it, *"our practical understanding of a given phenomenon is contingent on the data we choose to represent and measure it with."*

The ML research community has steadily introduced and explored new fairness definitions, and their associated metrics for responsible AI efforts (Narayanan, 2018; Barocas et al., 2019). A large portion of this work on fairness definitions and metrics for machine learning has focused on the domain of (binary) classification and regression. However, measuring fairness in recommendation systems poses distinct fairness challenges that are not shared with classification and/or regression systems (Ekstrand et al., 2022; Deldjoo et al., 2022; Kuhlman et al., 2021; Patro et al., 2022). This includes (but is not limited to): the systems' decisions being interdependent (ranked positions are rivalrous goods), occuring repeatedly over time, being personalized to users (introducing potential unfairness through user preferences), and involving subjective outcomes without a clear ground truth (Ekstrand et al., 2022). Additionally, recommender systems must balance competing fairness concerns among multiple stakeholders, such as recommendation consumers and item providers (Burke, 2017a). This has resulted in the creation of a specific area of research focusing on fairness within a recommendation or ranking context.

Here I outline some of the unique considerations of ML fairness for recommendation

systems in the form of "constraints" that impact decision-making during fairness operationalization.

The Multistakeholder Constraint

Recommendation system fairness is commonly referred to as **multistakeholder** fairness due to the goal of satisfying the fairness needs of multiple groups of stakeholders (Burke, 2017a). The two most common stakeholder groups are **providers** (those who provide or create content to be recommended) and **consumers** (those who interact with or consume the recommendations). Consumer-side or provider-side fairness can sometimes conflict with one another (Burke, 2017a). For example, prioritizing fairness for consumers by ensuring accurate and personalized recommendations may inadvertently disadvantage providers whose content does not align with those personalization goals, thereby creating a trade-off between the needs of the two groups.

The Group vs. Individual Constraint

Similar to measuring fairness in other domains of machine learning, evaluating fairness in recommendation systems could require differentiating between measuring group and individual fairness (Dwork et al., 2012). Recent work has highlighted that group and individual fairness both stem from similar principles, but differ when measuring them in practice (Binns, 2020). Group fairness measures if different groups of providers or consumers acquire similar recommendation outcomes, while individual fairness requires that similar individuals are treated similarly. Even further, for group fairness, some metrics seek to measure fairness for one group at a time (e.g., is a sensitive provider group's distribution of recommendations equal to their target distribution?), while others seek to measure fairness in a binary setting (e.g., comparing ranking distribution of sensitive

versus non-sensitive groups), or in a multi-group setting (e.g., what is the difference in distribution between providers from different countries in the world?) (Ekstrand et al., 2022). In recommendation and ranking, different metrics can measure group versus individual fairness within each stakeholder category; some individual metrics can also be adapted to measure different kinds of individual or group fairness (Ekstrand et al., 2022).

The System Component Constraint

Recommendation systems often consist of multiple components which leverage unique algorithms and sub-systems to create final recommendations for consumption. System components could include generating and retrieving pools of content, filtering said pools by ranking for high-priority goals, and finally re-ranking the final lists of content for consumption. This composite of system components results in a need to select at which step(s) to measure fairness, while also understanding how biases may change across the system (Liu et al., 2019; Sonboli et al., 2020c; Zehlike et al., 2017; Sonboli et al., 2020a).

Fairness could be measured over the entire population of recommended items (e.g., against a target distribution) or as a ranking problem (e.g., are the top-k rankings satisfying our fairness goals?). These types of choices can also have consequences for fairness evaluation results and the potential processing necessary to conduct such evaluations (Bower et al., 2022).

The Fairness Objective Constraint

Previous work has also attempted to categorize fairness metrics based on which fairness proxies are being measured. Verma et al. (2020) classified RecSys fairness metrics as *accuracy* based, *error* based, and *causal* based. The majority of fairness metrics were classified as accuracy based by these authors, where the metrics *"either state the condition*

for user satisfaction or provide a measure of deviation from the ideal ranking." Verma et al. (2020) also describe pairwise fairness metrics, introduced by Kuhlman et al. (2019), as error based metrics due to their measurement of false negatives and false positives against assumed ground-truth rankings.

Pairwise fairness metrics, inspired by previous binary classification fairness definitions, are designed to measure fairness objectives such as statistical parity or equality of opportunity. Causal based metrics require that item ranking is not related to group membership. More recently, Ekstrand et al. (2022) published an in-depth review of fairness in recommendation systems, introducing more nuanced ways to classify metrics beyond the three categories originally suggested by Verma et al. (2020). Notably, Ekstrand et al. (2022) categorized pairwise fairness metrics with accuracy metrics, alleviating the potential confusion between distinguishing when a metric measures error versus accuracy. The differences between these two publications reflect how fairness literature may change over time, making it difficult for practitioners to stay up-to-date and navigate this complex research space.

Related to this complexity, Raj and Ekstrand (2022) conducted an analysis of provider-side group fairness metrics and discovered that although many of these metrics shared basic fairness properties, in practice they could be optimized to measure different fairness objectives. For example, many group provider-side fairness metrics seek to measure weighted exposure of items, but various metrics might describe "fair weighted exposure" differently. If a practitioner chooses a fairness metric which requires that item exposure is related to item utility, they might measure fairness in terms of equal opportunity rather than statistical parity (Raj and Ekstrand, 2022). Both of these fairness objectives may be equally appropriate for a given fairness context, but a practitioner may not know which fairness objective(s) to optimize for or why—especially if they lack knowledge about the fairness goals of their organization.

It is clear that operationalizing fairness in recommender systems is a complex undertaking. The process of navigating from fairness goals to empirical measurements of those goals is something that machine learning engineers are not trained to do, and standards and guidelines in this domain are largely lacking. Fortunately, one approach from the social sciences has recently gained traction within the machine learning discipline, and might provide an optimal path forward: the method of measurement modeling.

2.2.3 Measurement Theory

> "The contested nature of fairness makes it inherently hard to measure: If there are multiple theoretical understandings of a construct, then it is imperative to articulate which understanding is being operationalized. A failure to do this makes it difficult to meaningfully compare different operationalizations because they may be operationalizing different theoretical understandings. In turn, this makes it difficult to identify mismatches between fairness as it is theoretically understood and any given operationalization," (Jacobs and Wallach, 2021).

Passi and Barocas (2019) describe that ML practitioners are challenged with the task of turning *"amorphous goals"* into *"well-specified problems"* that match their business objectives while also being measurable, predictable, and optimizable. As I noted in the previous section, the process of empirically translating goals into system design requires the use of proxies, or representatives of those goals that can be captured through observable data.

To specify what a particular human value means in the context of a real system is to **operationalize** that value (Stray et al., 2022). Corbett-Davies and Goel (2018) describe that it is notably difficult to operationalize the underlying definition of a proxy in a machine learning system, especially with respect to fairness proxies. This entire process

of scoping fairness goals, defining fairness based on those goals, selecting an appropriate fairness metric based on this definition, and incorporating it into the objectives of an ML system is what I refer to as **fairness operationalization** (Jacobs and Wallach, 2021; Smith et al., 2023a).

Now, since fairness is an inherently theoretical, subjective, and non-observable construct, it cannot be measured directly (Jacobs and Wallach, 2021)—which presents an opportunity to turn towards methods from other disciplines. In social science literature, **Measurement Theory** provides a framework for measuring unobservable, theoretical constructs, such as harm or fairness. Measurement theory has been used in previous work to operationalize user values in recommender systems (Milli et al., 2021), to create standards for measuring risk and impacts of generative AI systems (Chouldechova et al., 2024), and to confront the difficulty of appropriately modeling and measuring the contested concept of fairness (Jacobs and Wallach, 2021). For practitioners to operationalize fairness in a replicable fashion, they need to develop or have access to a cognitive model for determining what types of fairness metrics are most relevant in a given context—which is rarely straightforward (Smith et al., 2023a).

Stray et al. (2022) describe that *"the people involved in defining metrics have considerable influence over the ultimate function of a [machine learning system], which is why multi-stakeholder involvement in [ML] metric selection may be important."* However, this is often not the case in machine learning development—what is more common is that academics and practitioners develop and optimize for fairness metrics without input from other stakeholders. This lack of engagement with stakeholders while conducting fairness measurements is something that this dissertation seeks to confront.

In order to elicit a fitting metric for a human value, we can utilize what is called the **measurement modeling process** (Jacobs and Wallach, 2021; Box-Steffensmeier et al., 2008). A measurement model is a statistical model that maps unobservable, theoretical con-

structs to observable properties (Box-Steffensmeier et al., 2008). Although measurement models are the foundation for creating fairness metrics in ML systems, they are also susceptible to human bias. The process of selecting *which* observable properties are a best fit for an inherently unobservable property requires making many assumptions about causal relationships which may be impossible to empirically prove. When our assumptions are incorrect, a mismatch occurs between the theoretical construct we are attempting to measure and our proxy variable that we are using for that measurement—resulting in an unreliable metric. However, measurement theory also provides a set of methods that can be used to test the reliability of these metrics, and to search for potential mismatches.

Friedler et al. (2016) introduced the notion of a *construct space*—which is the space that captures the unobservable variables we would like to measure. These authors suggest that it would be good practice to explicitly state which assumptions are made about the relationships between constructs and observations when doing algorithmic fairness work (Friedler et al., 2016).

Two prominent approaches to increase transparency and scientific rigor around this construct space include improving (1) construct reliability; and (2) construct validity. **Construct reliability** is when similar inputs to a measurement model yield similar outputs—similar to statistical precision (Jacobs and Wallach, 2021; Box-Steffensmeier et al., 2008). **Construct validity** is when we can demonstrate that the measurements obtained from a measurement model are both meaningful and useful—similar to statistical unbiasedness (Jacobs and Wallach, 2021; Box-Steffensmeier et al., 2008).

Raji et al. (2021) also describe construct validity in machine learning as how well our research claims match our experiments. In the context of fairness, since one cannot "claim" a fully fair system, construct validity must capture the nuanced context and limitations of the measured fairness proxies.

> "Construct validity is an external validity issue related to how well an experimental setting relates to a research claim, or, in the context of machine learning, how well the benchmark dataset, and associated metrics of evaluation, represents a task. It concerns how well designed — or rather, how well constructed — the experimental setting is in relation to the research claim," (Raji et al., 2021).

To test for construct validity, practitioners can evaluate (1) face validity; (2) content validity; (3) convergent validity; (4) discriminant validity; (5) predictive validity; and (6) consequential validity. Each of these evaluations is described in detail in Table 2.1.

Previous work also describes that good measurements can be evaluated based on reliability and validity (Stray et al., 2022; Quinn et al., 2010). *"The evaluation of any measurement is generally based on its reliability (can it be repeated?) and validity (is it right?). Embedded within the complex notion of validity are interpretation (what does it mean?) and application (does it "work"?)"* (Quinn et al., 2010).

Another benefit of using measurement theory to help operationalize fairness is that this method has been proven to work in real-world contexts as well as theoretical contexts. Milli et al. (2021) used measurement theory on a real-world recommender system to measure "value" for users on the Twitter platform. This research proved that measurement theory can be useful in practice for ML practitioners who are struggling to appropriately measure theoretical constructs in a real system.

Challenges with Measurement Theory

As likely expected, a complex methodology like measurement theory does not come without complications and challenges. Stray et al. (2022) raise an important challenge when

Table 2.1: Steps to evaluating construct validity in machine learning models, adapted from Jacobs and Wallach (2021).

Steps	**Description**	**Probes to Evaluate**
Face Validity	Subjective first pass to check if the measurements obtained from a measurement model look reasonable.	Do the measurements look like a reasonable depiction of the theoretical construct?
Content Validity	A check to see how well the measurement captures the theoretical construct (or which interpretation of that construct is being measured).	Do the measurements capture the specific definition of fairness we have chosen? Do our proxies capture only observable variables related to the original construct and nothing more? Does our measurement capture the relationship between the non-observable construct and its proxy variables?
Convergent Validity	A check to see how much the current measurement differs from previous measurements of a similar construct.	Do the measurements give the same results as a previous metric (if so, is this new metric necessary?) Do the measurements give different results as a previous metric (if so, are these differences justified?) Do the measurements give subtly different results as previous metrics (if so, are those differences justified?)
Discriminant Validity	A check to see how this measurement might unintentionally measure other constructs.	Are there any correlations between our measurements and measurements of other constructs that are not related to our fairness definition?
Predictive Validity	A check to see how useful the measurements are.	How well do these measurements predict observable / non-observable properties related to our theoretical construct?
Consequential Validity	A check to acknowledge the consequences of this measurement model.	How is the world shaped by these measurements? What world do we wish to live in?

attempting to measure theoretical constructs:

> "Even a good metric will change meaning when it is known to be used to make consequential decisions, e.g. student test scores must be interpreted differently when they are used to decide academic progression, because instructors will begin "teaching to the test". This effect is sometimes known as Goodhart's law, but there are a variety of different causal structures which can produce feedback processes that widen the distance between a metric and the underlying value" (Stray et al., 2022).

Additionally, certain constructs are inherently harder to measure than others (Stray et al., 2022; Jacobs and Wallach, 2021). When a theoretical construct has many (possibly competing) definitions, it is necessary to select *one* definition to measure at a time. This implies that when attempting to measure fairness, practitioners will need to select one single definition of fairness to measure at a time, which involves making value tradeoffs that might benefit certain stakeholders over others. I further explore the implications of fairness definitions and how different definitions might differentially impact different stakeholders during my analysis of Study #3, in Section 5.2.

As Selbst et al. (2019) describe, ***"there is no way to arbitrate between irreconcilably conflicting definitions [of fairness] using purely mathematical means."***

Mehrotra et al. (2017) found that another challenge with measurement in ML systems is that different demographics may interact with the system in different ways, which could skew metrics even after controlling for these differences. Stray et al. (2022) describe this challenge as its own fairness problem. *"For example, if older users read more slowly than younger users, then a metric based on dwell time will be an over-optimistic measurement for older users regardless of their level of satisfaction. Thus, interpreting metrics at face*

value may systematically disadvantage and misrepresent certain demographics and user groups," (Stray et al., 2022).

One might argue that the challenges of measuring "immeasurable" things (such as theoretical constructs) raises the question of whether it is worth attempting to measure them in the first place. The assumptions that must be made to accurately reflect an unobservable construct through an observable proxy necessarily introduce opportunities for bias and limit our ability to appropriately measure and optimize ML systems for these constraints.

For example, in the context of image cropping on Twitter, Yee et al. (2021) describe that sometimes our assumptions are simply incorrect, and in these situations, reducing theoretical concepts to numbers might mean that our measurements will never fully represent reality.

> "Framing concerns about automated image cropping purely in terms of demographic parity fails to question the normative assumption of saliency-based cropping: the notion that for any image, there is a "best"–or at the very least "acceptable"–crop that can be predicted based on human eye tracking data. This critique echos previous calls to question machine learning's imposition of a normative "optimal state" in all aspects of life, especially those related to human expression and the arts" (Yee et al., 2021).

However, therein lies the crux of doing what I call **pragmatic ethics**—when there is no "correct" solution, and yet a decision must be made. In Philosophy courses, it is useful to debate the various approaches of defining and measuring fairness, discussing the implications of each choice, and recognizing that no choice is necessarily "wrong" nor "right." These philosophical debates also often lead to discussions about whether there exists a universal conception of "good" or "bad."

"Technology is neither good nor bad; nor is it neutral..." (Kranzberg, 1986)

However, in machine learning technologies, these debates, while also useful, must result in actual *decisions* for how to design the system. Now, there is also a point to be made that deciding to not make or deploy the technology in the first place is a legitimate decision in the event that no appropriate metric exists to evaluate for constructs such as fairness, when unfairness could pose too high a risk on stakeholders (Baumer and Silberman, 2011).

Realistically, more often than not, the decision to "not design" is not an option for practitioners. In these scenarios, to do pragmatic ethics work is not to find the "correct solution," but instead it is to recognize that there may be no "solution" at all. Instead, pragmatic ethics asks practitioners to put in the work that it takes to understand the consequences of various approaches to integrating ethics into the system (e.g., through measuring and optimizing for fairness), to make a decision about which approach is best (e.g., according to the goals and values of their organization and its stakeholders), and to take informed responsibility for the consequences of that approach.

Or, as Stray et al. (2022) describe: *"at some point, a real recommender must commit to specific operationalizations of broad concepts, with the resulting tradeoffs between competing values and stakeholders."*

2.2.4 Designing for Human Values

If the goal of operationalizing fairness for recommender systems is to design recommender systems to "be more fair," the work must extend beyond measurement and optimization. Recommender systems used in online platforms are highly complex, created from a combination of algorithmic elements including feature extraction, ranking, flagging, re-ranking, scoring, and/or clustering using a variety of technologies and various forms of

machine learning. Recommendation systems built from interactions between large, complex sub-systems can be difficult to understand even for those who create them. Thus, incorporating fairness into a recommender system must take into consideration the design of the entire system, and how this system might interact with its various stakeholders.

Knobel and Bowker (2011) describe **cultural valence** in technology design as the recognition that different people hold different values, and because of this, designers of technology might have different values than those of the stakeholders of their technology. Thus, it is imperative that when incorporating values into technology design, robust user-centered and human-centered methodologies are used.

Human-centered design is a framework that seeks to center human rights, needs, creativity, and dignity at the center of the technological design process (Yee et al., 2021; Buchanan, 2001). Foundational to human-centered design is the concept that humans both shape and are shaped by our technologies, which invites the opportunity to design the kind of world we want to live in (Gasson, 2003). While user-centered design focuses on solving human problems through technological design, human-centered design emphasizes the non-technical facets of a problem that might impact the design of a technical intervention (Gasson, 2003). For example, human-centered design might consider the social, cultural, and ethical factors, such as the values, power dynamics, and broader societal implications that influence both the problem space and the potential outcomes of a technical intervention. In recent years, the framework of **Design Justice** has provided robust methods for engaging users and their communities in technological design in a way that captures a community's values and the various potentially intersectional identities that exist among its members (Costanza-Chock, 2020).

A **value** can be defined as something that a person or group of people consider to be important in life (Borning and Muller, 2012). One methodological approach for integrating specific values into the design of a technology is through **Value Sensitive Design (VSD)**,

originally developed by Batya Friedman (Friedman et al., 2002). VSD is an iterative design methodology that *"integrates conceptual, empirical, and technical investigations,"* (Friedman et al., 2002).

In VSD, a **conceptual investigation** is a philosophically informed analysis of the central constructs and issues under investigation, while **empirical** and **technical investigations** explore how humans and technology can be observed and measured for their ability to support or hinder human values. Using measurement theory, developing measurement models, and incorporating fairness metrics into an ML system are all examples of technical investigations. Conversely, conceptual investigations are philosophically informed analyses of the constructs being investigated (Friedman et al., 2002). As such, measuring and optimizing for fairness is only one part of incorporating the value of fairness into ML design. Another equally important part of the process is to conduct philosophically informed analysis of fairness and to utilize this knowledge to help inform the design of our technical investigations.

I now turn towards conceptual investigations to guide in the design of fairness evaluations. In the following section, I describe the most popular ethical theories from moral philosophy, and how they can be utilized to inform the design of fairness operationalization in machine learning systems.

2.2.5 Ethical Theories

Ethics can be understood through several interrelated categories. **Meta-ethics** explores the nature, origins, and meaning of ethical principles, asking foundational questions such as "What is morality?" and "How can we know what is right or wrong?" **Descriptive ethics** examines how people actually behave and make moral decisions, focusing on cultural, societal, and individual practices without prescribing what should be done. **Normative**

ethics, by contrast, seeks to establish frameworks for determining what people ought to do, offering principles or rules for guiding moral behavior. Finally, **applied ethics** involves the practical application of ethical principles to specific issues or contexts, such as medicine, technology, or business.

Relational and rational ethical theories both fit within the domain of normative ethics, as they propose frameworks for assessing moral rightness or wrongness. **Rational ethical theories** have historically dominated western philosophical frameworks and our notions of what is morally "right" or "wrong". Rational ethical theories focus on rules, principles, and reason to determine right and wrong. Conversely, **relational ethics** is a more collectivist approach to morality, and emphasizes the importance of relationships when determining what is right and wrong. Relational ethics are more concerned with the situational context and relationships involved in a decision, rather than universal principles that must be followed. By focusing on relationships, relational ethics challenges the universality of rational theories and highlights the dynamic, interpersonal nature of ethical decisions.

In the context of fairness, both of these approaches could be equally appropriate when defining what fair and unfair treatment might look like from an algorithmic system. However, most literature on ML fairness has taken a rational ethics approach, because principles and rules are easier to quantify and scale when compared to relationships and social contexts (Zoshak and Dew, 2021; Birhane, 2021). Here I describe some of the dominating ethical theories that could be applied when operationalizing fairness in machine learning, categorized by their rational or relational roots.

Rational Ethics Theories

Consequence-based fairness has roots in Jeremy Bentham's theory of *Act Utilitarianism*, and posits that *"an act is right if and only if it results in at least as much overall well-*

being as any act the agent could have performed" (Eggleston, 2014). For an act utilitarian system, optimizing fairness is synonymous with maximizing positive outcomes and minimizing negative outcomes. This operationalization of fairness is common in computer science research, as it offers two relatively straightforward methods for measuring and optimizing fairness: (1) select a proxy for "positive outcome" and tune the system to maximize that metric; or (2) select a proxy for "negative outcome" and tune the system to minimize that metric (e.g., (Karnouskos, 2021)). However, optimizing an ML system for act utilitarianism might maximize benefits for all users in aggregate, while failing to maximize benefits for subgroups of users that fall within protected demographic categories such as race or gender. An example of this issue is the design of facial recognition technologies that did not correctly identify racial minorities (Buolamwini and Gebru, 2018).

Contract-based fairness is rooted in *Social Contract Theory* (Laskar, 2013), which involves defining an ideal social contract and abiding by its rules. One interpretation of contract-based fairness is Rawls' theory of *Distributive Justice* (Gabriel, 2022), which seeks to equally distribute resources within a society. One way to advance distributive justice is through *Equality of Opportunity* (Arneson, 2002), which requires that everyone has a fair chance to receive benefits. Equality of opportunity has an extensive history in financial regulation (Barocas et al., 2017), as there are global, systemic inequalities among opportunities to obtain wealth (Sawyer and Jarrahi, 2014). One way that inequality of opportunity arises in recommendation is through *popularity bias*, when recommendation algorithms exacerbate the difference in exposure between items at different levels of popularity, also known as the long-tail effect. There is a substantial research literature that seeks to mitigate the effects of popularity bias in recommender systems (see e.g., (Park and Tuzhilin, 2008; Channamsetty and Ekstrand, 2017; Abdollahpouri et al., 2019; Wei et al., 2021; Zhu et al., 2021; Yalcin and Bilge, 2022; Burke et al., 2022b).

Duty-based fairness is rooted in Kant's theory of *Deontology* (Larry and Moore, 2016).

Achieving fairness through deontological principles means consistently following a set of rules. Optimizing for deontological fairness in a machine learning system requires selecting a set of rules (e.g., model or algorithm heuristics) and then consistently following those rules. Ferraro et al. (2021) provide an example of duty-based fairness as it relates to music recommendation on Spotify. One of their proposed platform "duties" is a promise to musicians (item providers) that less popular or new artists will be recommended at a fair rate in comparison to more popular artists.

Relational Ethics Theories

Character-based fairness is rooted in Aristotle's theory of *Virtue Ethics* (Hursthouse, 2007). Virtue ethics holds that decisions or outcomes should be based on a person's *character* rather than their *actions*. However, in recommendation, users' actions are often used as proxies for one's "character," (e.g., when a users' interests and personality are inferred from their engagement (Steck, 2018)). One example of character-based fairness in recommender systems is the design of trust-based recommender systems, which serve recommendations from individuals that users have chosen to trust, in contrast with collaborative recommendation in which the peers influencing the recommendation are unknown (Victor et al., 2011; Li et al., 2020).

Ubuntu-based fairness is rooted in several African philosophical traditions and emphasizes the importance of context in ethical decision-making (Metz, 2011). Ubuntu ethics recognizes that ethical behavior is influenced by the specific cultural, social, and historical context that individuals and communities are situated in, and focuses on understanding and respecting these cultures and contexts. Beyond respect, this form of fairness centers the values of empathy, generosity, and compassion (Gwagwa et al., 2022). To the best of my knowledge, previous work has not yet explored how ubuntu-based fairness could impact ML fairness operationalization. However, one example of how ubuntu-based fairness

could be incorporated into the design of machine learning systems could involve optimizing a system for multiple fairness considerations at once, and dynamically allocating different fairness interventions based on the geographic and cultural context of the users involved in a specific decision, such as through social choice theory (Burke et al., 2020).

2.3 Applying ML Fairness in Industry Settings

> "As human-computer interaction researchers, we often make arguments to stakeholders about how and why they can change technology to better serve users and/or improve society. In the case of algorithmic fairness, stakeholders such as regulators, lawmakers, the press, industry practitioners, and many others have the opportunity to take positive action. Technology companies in particular have tremendous leverage to improve algorithmic fairness because they are immediately proximate to many of the technical issues that arise, and they are uniquely positioned to diagnose and develop effective solutions to complex problems that would be difficult for outsiders to address," (Woodruff et al., 2018).

Despite the resurgence of interest in incorporating fairness into ML design, doing this work in practice is still quite challenging. As described in section 2.2.2, the process that practitioners face when scoping fairness goals and selecting an appropriate fairness metric for a recommender system is not straightforward nor simple. This process includes but is not limited to the selection of: which stakeholders are most relevant for a fairness intervention (including which subpopulations of those stakeholders, and whether fairness will be measured for groups or individuals within those subpopulations); which system components will be evaluated for fairness; which data needs to be collected or annotated to conduct an evaluation; which definition(s) or fairness goals are most relevant for the given

scenario; which proxies are most appropriate for a given evaluation; and which metric(s) are most aligned with the given fairness goals and available fairness proxies (Smith et al., 2023a). Each of these steps is quite complex, and involves extensive research or expertise on ML fairness topics, as well as interdisciplinary and cross-team collaboration.

Fortunately, a lot of research on ML fairness has already been conducted, which provides many options for practitioners to select from when deciding which fairness metrics to use during a fairness evaluation. As I described previously, when incorporating fairness goals into ML systems, this involves determining a set of criteria or requirements that must be met in order for the system to be deemed "fair" under a specific definition (Smith et al., 2023a). ML fairness definitions have been introduced most often for classification (e.g., equal odds (Hardt et al., 2016), treatment equality (Berk et al., 2021), equal opportunity (Hardt et al., 2016; Agarwal et al., 2018), statistical parity (Dwork et al., 2012), error parity (Buolamwini and Gebru, 2018), false positive and false negative parity (Zafar et al., 2017; Hardt et al., 2016), and omission rate parity (Zafar et al., 2017; Kleinberg et al., 2016)) and more recently for recommendation (e.g., search neutrality (Grimmelmann, 2010), top-k fairness (Zehlike et al., 2017), calibration parity and classification parity (Singh and Joachims, 2018)).

Fairness definitions and interventions have also been incorporated into real-world recommendation contexts including (but not limited to) microlending recommendation (Liu et al., 2019), job recommendation (Geyik et al., 2019), and music recommendation (Beattie et al., 2022).

However, despite the proliferation of research papers that approach ML fairness, practitioners still face many challenges when doing fairness work. Considering the quantity and complexity of fairness definitions, different definitions will likely be put forward by different stakeholders, all of which must be integrated within a practitioner's fairness intervention. Beyond the practical challenges practitioners face when selecting appropriate fairness def-

initions and goals, other barriers exist that can make it challenging for practitioners to operationalize fairness. Rakova et al. (2021) previously noted that apart from practical challenges in implementation, *"[responsible] AI initiatives also require operationalization within — or around — existing corporate structures and organizational change."* Integrating fairness work into existing corporate or business structures can be a difficult task, especially considering trade-offs that may occur (e.g., if optimizing fairness for one set of stakeholders diminishes fairness for another set of stakeholders (Burke, 2017a)). Ensuring that fairness interventions are in alignment with existing business goals, and resolving perceived tensions against fairness work is crucial to responsible design efforts (Rakova et al., 2021).

One example of a tension previously introduced in the literature discusses the gap between academic research and practical implementation (Holstein et al., 2019). Specifically, this research highlights that domain-specific, scalable standards would be greatly beneficial, (as also discussed by (Green and Hu, 2018; Lee, 2018)). However, other research has objected to scaling, standardizing, or automating ML fairness in practice, due to the context-dependent nature of fairness work (Lee et al., 2020; Wachter et al., 2021).

Researchers and practitioners have turned towards tooling to combat some of these challenges. **Tooling** in the context of this work includes open-source libraries that provide code to implement fairness metrics, as well as protocols, questionnaires, and worksheets to help educate and support decision-making in an algorithmic responsibility setting (e.g., (Madaio et al., 2020; Cramer et al., 2018a; Moss et al., 2021; Bellamy et al., 2019; Xu and Doshi, 2019; Bird et al., 2020; Richardson et al., 2021; Metcalf et al., 2021)). However, despite all of these available tools, practitioners still report a disconnect between these tooling efforts and what they actually need in practice (Madaio et al., 2022; Law et al., 2020; Veale and Binns, 2017; Lee and Singh, 2021; Richardson et al., 2021).

One major critique of these toolkits is that they exclude the everyday practices and

decision-making processes that are a critical part of doing responsible technology design.

> "In instantiating ethics through reproducible processes like checklists or programming packages, these efforts may manage only to offer the illusion of completion, and in doing so imply that ethics "has been done." Rather, it is precisely in these everyday practices that meaningful ethics are located and robust anticipation of social harms could be contemplated" (Metcalf et al., 2019).

Another major critique of ML fairness tools is that they do not provide the necessary guidance for practitioners to *begin* exploring their fairness goals (Lee and Singh, 2021). Open-source libraries that provide metrics for practitioners are of little use if the practitioners do not know how to define the impact they want to measure, or scope their measurement context.

One recent study explored how much agency AI developers report experiencing in their work with respect to ethics, and discovered that there is no consensus over how much responsibility practitioners have over the ethical implications of their technologies (Griffin et al., 2023). Even in the event that all practitioners felt a moral obligation to incorporate fairness into the design of their technologies, most ML practitioners are not trained in disciplines like ethics or philosophy (Saltz et al., 2019), which creates another barrier to entry for this complex decision-making space.

Tasks such as scoping fairness goals or deciding on appropriate fairness constraints might seem very daunting without institutional or academic knowledge of ethical theories or fairness metrics.

Passi and Barocas (2019) detail this challenge of scoping as fairness "problem formulation," and describe how different formalizations of the same fairness problem can lead to different

operationalizations of fairness goals that relate to that problem. The authors emphasize the importance of the social and historical context that a fairness problem may arise within, and how different stakeholders might be impacted by this problem and its proposed solutions (Passi and Barocas, 2019).

In response to some of these challenges with formalizing fairness problems, goals and considerations, Saleiro et al. (2018) created a decision tree to help practitioners select an appropriate ML fairness operationalization for their given context. However, this decision tree assumes that the practitioner has prior knowledge of policy and ethics jargon, with some branches in the tree asking questions like, *"are your interventions punitive or assistive."* Additionally, this decision tree was designed for the context of binary classification, not ranking or recommendations. In the context of recommendation systems, fairness metrics and considerations are vastly different from a binary classification setting, especially since outcomes are not necessarily binary nor measurably favorable. There is rarely a "ground-truth" to compare the final recommendation lists against beyond assuming user engagement as a positive prediction (Beutel et al., 2019b).

All of these tensions and challenges that I have raised in this chapter hinder practitioners' abilities to operationalize fairness in practice. Without proper guidance for practitioners to scope their fairness goals and to select metrics that appropriately map to those goals, fairness evaluations in practice could hinder or even diminish fairness at a large scale (Corbett-Davies and Goel, 2018).

In addition, the practical constraints of industry requires that fairness operationalization does not deter from business goals, which can drastically impact both the goals and the design of fairness evaluations (Rakova et al., 2021).

Finally, while these practical challenges of implementing fairness in real-world systems are substantial, they underscore the importance of continuing this work. In this dissertation,

I tackle the multifaceted complexities of operationalizing fairness in real-world recommendation systems. By examining the experiences and preferences of impacted stakeholders alongside the constraints faced by practitioners, I explore potential best practices for advancing fairness in machine learning applications. This work offers actionable recommendations for navigating these challenges and lays a foundation for ongoing efforts to improve fairness in dynamic, real-world contexts.

Chapter 3

Methods

In this chapter, I outline all of the methods for each of the five studies included in this dissertation. Four of the studies were semi-structured interview studies (Studies # 1, 3, 4, and 5), while Study #2 was a focus group study. Here I describe the specific methods for each study in the order they appear in this dissertation.

3.1 Study #1: The End-Users' Perspective

Collaborator Statement: For this study, I had the privilege of collaborating with Nasim Sonboli, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. My contributions to the research included conducting multiple interviews, leading the analysis and coding of all interview transcripts, and working closely with my co-authors to co-write the resulting paper.

In this work, my collaborators and I used interviews as a way to explore the stories and experiences of a sample of users of recommender systems. Though some previous work has stated that user input for algorithmic design may not be helpful for engineers, due to

users' potential lack of knowledge of the system (Saxena et al., 2019), I argue that it is precisely this lack of knowledge that can be beneficial for the design and deployment of *explanations* in systems. Understanding what the users do *not* know about a system can help the engineer pinpoint where users might need more education about the platforms they are interacting with, and how an explanation can meet those needs.

As an initial exploratory study, we recruited a convenience sample of 30 undergraduate and graduate students from the University of Colorado Boulder, all of whom had prior exposure to recommender systems (e.g., on e-commerce or streaming platforms such as Amazon or Netflix). Our participants included 18 women and 12 men, with an age range of 18 to 32. 15 participants were studying Computer Science, Information Science, or related fields, and 15 were pursuing other majors. Participants were recruited via social media postings, a university announcement board, and advertisements in classes, and were compensated $10 for their time; this study was approved by the University of Colorado's Institutional Review Board. Though our sample of young internet users provides a good foundation for understanding opinions and concerns that ordinary recommendation consumers might have about fairness, I recognize that it is limited in scope and I do not suggest that our findings are generalizable to a broader population.

My collaborators and I conducted face-to-face, semi-structured interviews with participants that began with a discussion of their past experiences with recommender systems and their folk theories of recommendation algorithms and fairness objectives. We asked them to explain their current knowledge about how these systems work, and what fairness meant to them in this context. For the latter half of the interviews, we used the Kiva crowd-sourced microlending platform (explained in more detail in Section 3.1.1) as a case study to explore the ways that explanations can help or harm the participant's understanding of a fairness-aware system. We used Kiva as a case study both to have a concrete example to ensure consistency across participants, and because Kiva is a real-world ex-

ample that contains fairness concerns. In our interviews, we asked the participants to consider what helpful explanations might look like in Kiva's platform if fairness-aware recommendation were implemented.

Following interview transcription, my collaborators and I conducted a version of thematic analysis (Braun and Clarke, 2006), where we conducted open-coding on a subset of the interviews using MAXQDA software, and then met to confer on, synthesize, and finalize a set of themes that best captured the insights gathered from our participants.

3.1.1 Kiva as a Case Study

Kiva is an organization seeking to enhance financial inclusion in the world through microlending. Their platform provides a space for those who are seeking loans (borrowers) to get funded by those who are seeking to supply capital (lenders). Kiva's mission emphasizes equitable access to capital for all borrowers, who are individuals without access to traditional forms of capital, to improve their living situations. In the recommendation ecosystem of Kiva, the providers of recommended content are the borrowers, and the consumers of recommendations are the lenders (Choo et al., 2014). Naturally, a recommender system in this context raises an important fairness concern. If borrowers are recommended inequitably, the system could be contributing to an inequitable pattern of resource distribution on its platform. However, if fairness considerations make the recommendation ecosystem unsatisfactory for lenders, causing them to leave the platform, this could cause negative consequences for borrowers as well.

3.2 Study #2: The Providers' Perspective

Collaborator Statement: For this study, I had the privilege of collaborating with Aishwarya Satwani, Robin Burke, and Casey Fiesler. My contributions to the research included designing and facilitating all focus groups, leading the analysis and coding of all focus group data, and leading the writing for the resulting paper.

This study consisted of four virtual focus groups with a total of thirteen participants, which is in line with standards for user studies in HCI (Caine, 2016). Focus groups were the most appropriate method to adopt for this work because of our goals to uncover tensions between providers and to encourage brainstorming that extended beyond participants' individual experiences through group discussions. In addition, focus groups encourage research participants to develop ideas collectively, and sharing experiences among participants can prompt people to elaborate on their stories and themes, which can help researchers interpret those experiences (Smithson, 2008). The first two focus groups included participants who are content creators on platforms like TikTok, Instagram, or YouTube. The final two focus groups included participants who use dating apps. This research was approved by the University of Colorado's Institutional Review Board.

3.2.1 Participants

I recruited participants via open calls to participate on Twitter, LinkedIn, Instagram, TikTok, and Slack channels with personal networks. Interested participants were sent to an online form to select which type of provider they were. All participants were compensated $30 USD for their participation. Since the goal of this study was to elicit the perspectives and preferences of recommendation *providers*, I asked participants to concentrate on this role during the focus groups. For example, even though content

creators become consumers when they scroll others' posts or dating app users become consumers when they swipe on others' profiles—both of these groups are also providers (users whose content/profile is exposed to others). To ensure I elicited preferences of the provider, I asked participants to engage in activities while remembering their vested interest in getting their content or profile exposed to other users. I chose to focus on providers because their perspective is often missing from recommender system design and evaluation and also because, especially in the case of platforms where providers can be paid, there are higher stakes for providers.

All participants and their collected attributes are included in Table 3.1. I included gender identity and race/ethnicity in the tables as self-described by participants, to preserve their preferred terminology. I note the lack of diversity in both gender identity and racial/ethnic identity of our recruited participants as a limitation of this research study, and the results should be interpreted with this in mind. The majority female, majority White demographic of our participants likely omits useful dimensions of unfairness experienced by other demographic groups, but still provides a helpful starting point for this seminal work. Each participant was assigned an alias based on their recommendation domain. For example, content creators were assigned the alias PC#, whereas dating app users were assigned the alias PD#. (e.g., "PC1" or "PD1"). I use these aliases as reference throughout Section 4.3. Through the pilots, I found that limiting focus groups to 3-4 participants was ideal to allow for active participation.

3.2.2 Focus Group Design

The focus group design was the same for all four focus groups, regardless of the recommendation domain in question. All participants were guided to an online collaboration document on Google Docs. For reference, the collaboration document used for content

Table 3.1: Demographics of content creators. FG# refers to the associated focus group.

Alias	**FG#**	**Type of Content**	**Platform(s) Used**	**Follower Count**	**Age Range**	**Gender Identity**
PC1	1	Art & sewing	Instagram, YouTube	500-1000	20-30	Female
PC2	1	Short/ long-form animation	LinkedIn, Vimeo, Tumblr, Instagram, YouTube	<500	20-30	Nonbinary
PC3	1	Historical education	YouTube, Instagram, TikTok, Facebook, Threads, Patreon	>10,000	30-40	Female
PC4	2	Movement videos	Instagram, TikTok, Facebook	>10,000	20-30	Female
PC5	2	AI-facilitated digital art	TikTok, Instagram	<500	50-60	Male
PC6	2	Artist/ Illustrator	Instagram, Facebook	4,000-10,000	40-50	Female

creators and the associated focus group protocol can be found in the Appendix in Section 7.4. Each activity involved independent brainstorming in the shared document, as well as a group discussion afterward. All focus groups were conducted virtually via Zoom and were recorded on the same platform. Audio recordings were then transcribed using Microsoft Word's audio transcription software. The focus groups included 4 different activities for participants:

1. **Introduction Activity.** Participants brainstormed about their experiences with the recommender systems that they are a provider for, as well as their perception of the algorithms on these platforms. Then all participants introduced themselves and shared with one another.

2. **Unfairness Activity.** Participants brainstormed about their experiences of unfair-

Table 3.2: Demographics of dating app users. FG# refers to the associated focus group.

Alias	**FG#**	**Dating Apps Used**	**Paid for Premium?**	**Age Range**	**Gender Identity**
PD1	3	Tinder, Bumble, OkCupid, Hinge, Taimi, Her, Lex, Grindr, Feeld, Facebook Dating	No	20-30	Female
PD2	3	Hinge, Tinder	Yes	20-30	Male
PD3	3	Hinge, Tinder, Bumble	Yes	20-30	Nonbinary/ Agender
PD4	4	Tinder, Bumble	No	20-30	Female
PD5	4	Hinge, Tinder, Feeld	No	20-30	Female
PD6	4	Hinge, Bumble, Tinder, The League, Coffee Meets Bagel	Yes	20-30	Female
PD7	4	Bumble, Hinge, Feeld, Tinder	No	20-30	Cisgender Woman

ness and fairness as providers. Participants were not provided with a definition of "fairness" or "unfairness," and were asked to recall their experiences based on their personal perceptions of what fair treatment was for them. After brainstorming, participants volunteered to discuss their experiences, while the research team took notes to come up with a "fairness concerns list" that captured these experiences.

3. **Fairness Goals and Definitions Activity.** Participants collectively selected a recommendation platform in their domain that everyone was familiar with (e.g., Hinge for dating app users, or Instagram for content creators). They brainstormed a list of fairness goals related to this platform, with specific attention towards the fairness concerns list that was created during the previous activity. Participants also developed a condensed fairness definition that attempted to capture these fairness goals. Together, the group discussed their fairness goals and definitions, as well as their challenges and experiences with the activity.

4. **Fairness Metrics Activity.** Using the same platform as decided in the previous activity, all participants brainstormed at least one "Fairness Metric" that allowed them to measure whether or not one of their fairness goals were being achieved by the recommendation system. They were asked to include details about (1) which data would need to be collected to measure fairness in this way; (2) which user populations (demographics) might need to be compared against one another; (3) how they would know if "fairness" had been achieved; and (4) how they would know if the platform was still not "fair" enough. Together, the group discussed their experiences with trying to design fairness metrics, including identifying any tradeoffs and challenges with their specific metric(s), and how they would like this process to be done in practice.

3.2.3 Data Analysis

Using the transcribed audio recordings of the focus groups, I conducted a version of thematic analysis via inductive open coding, by tagging individual sections and quotes with labels that were associated with that particular observation. Written responses from participants during brainstorming sessions were also coded. When a new response or quote fit into a previously defined code, it was included in that category. After creating a codebook with related quotes and observations, I discussed the observations and codes with the rest of the research team and came to a consensus about emerging themes. These high-level themes are discussed throughout Section 4.3.

3.3 Study #3: The Practitioners' Perspective (Kiva)

Collaborator Statement: For this study, I had the privilege of collaborating with Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. My contributions to the research included leading the design of the the four fairness logics, leading the analysis and coding of interview transcripts in relation to these fairness logics, and working closely with my co-authors to co-write the resulting paper.

My collaborators conducted semi-structured interviews with 23 employees of Kiva's microlending platform. They transcribed the interviews, and then we worked together to analyze the transcripts using thematic analysis (Braun and Clarke, 2006), resulting in a framework of fairness logics and characterizations of tensions among those logics. This research was approved by the University of Colorado's Institutional Review Board.

3.3.1 Participants

My collaborators recruited 23 Kiva employees, each with a job role that had an investment in the recommendations of loans posted on Kiva's website. The participants worked from five countries across three continents. They worked for nine different teams within Kiva. Recruitment of initial participants was based on the advice of and introductions from Pradeep Ragothaman, who is Head of Data Science at Kiva. Subsequent recruitment was done through snowball sampling. Due to the relatively small size of some of the teams at Kiva and the number of uniquely identifiable job roles, I report participants in this section in aggregate in order to support their anonymity. In what follows, I refer to the participants numerically as P1 through P23. Transcripts from P6 and P10 are not included in this analysis as they were pilot interviews for a future study with a Kiva lender and with an employee of one of Kiva's lending partners.

3.3.2 Data Collection

My collaborators conducted the interviews in this study, they tailored interview questions to each individual and their specific role at Kiva, with a focus on the following themes: their role at Kiva; how they apply the value of fairness in their work; challenges they experience when applying the value of fairness in their work; and how they interact with different technologies at Kiva. Interviews were conducted and audio-recorded via Zoom video conferencing.

3.3.3 Data Analysis

I, alongside my collaborators, conducted collaborative data analysis via the best practices of thematic analysis (Braun and Clarke, 2006). After all interviews were completed, we began conducting open coding on interview transcripts; a generative process in which two authors independently tagged low-level themes in the transcript. The first round of open coding resulted in categories that included different definitions of fairness (e.g., fairness is defined as maximizing impact for borrowers), different ways that fairness was operationalized through work (e.g., determining impact scores, auditing for fairness concerns), and different fairness heuristics that focused on considerations for different groups of stakeholders (e.g., prioritizing risk over impact, educating lenders or borrowers). We then turned to explore the relationships among codes. In our next round of coding, we attempted to organize our data by stakeholder, building on related research (Lee et al., 2019b) that has successfully used multistakeholder analysis to understand co-occurring fairness concerns. This round of coding ultimately felt insufficient in organizing the data as the relationship among stakeholders and fairness concerns was neither one-to-one nor clear cut.

In our second iteration, I switched to organizing our data by the underlying logic of the fairness consideration. This phase of analysis occurred in parallel with a substantive analysis of the fairness research literature to identify which fairness constructs and preexisting logics were most related to participants' descriptions of fairness at Kiva. The results of this coding phase were also used to organize the literature review for this paper. Our next round of coding, then, switched to being more deductive, using the classes of fairness from the literature review to organize the data. Following this analysis, key excerpts of interviews remained that did not fit neatly into any one class of fairness. Instead, they hinged on conflicts between or intersections among classes of fairness. Our final analytic phase focused on classifying the remaining excerpts in terms of these intersections.

3.4 Study #4: The Practitioners' Perspective (Spotify)

Collaborator Statement: For this study, I had the privilege of collaborating with Lex Beattie and Henriette Cramer while interning for the Algorithmic Impact Team at Spotify. My contributions to the research included conducting the literature review that led to the initial design of the decision tree, leading the design of the decision tree, designing and conducting all interviews, leading the analysis and coding of interview results, and leading the writing for the resulting paper.

To capture the complex challenges practitioners face when scoping and identifying fairness metrics, my collaborators and I iteratively designed a decision-making framework based on a literature review and feedback from our participants. Our framework is a decision tree designed to specifically scope quantifiable harms and corresponding metrics from a pre-defined, conceptual, potential harm of the system. The decision tree was created to help practitioners decide between various fairness constraints to scope which quantitative fairness context and metric category is most appropriate for their goals. The final iteration

of the decision tree can be found in the Appendix in Section 7.4.

I created the initial design of this decision tree to mirror past frameworks introduced by Aequitas and Fairlearn, which helped scope quantitative fairness contexts and identify fairness metrics for binary classification and regression (Saleiro et al., 2018; Bird et al., 2020). To the best of my knowledge, I was unable to find versions of these types of tools for recommendation or ranking systems.

We then used this decision tree framework and an associated spreadsheet of fairness metrics during interviews with ML practitioners at Spotify. To glean more nuanced observations from interviews, my collaborators and I iterated on the framework to account for participant feedback. Our iterative design process was inspired by previous research on co-designing ML fairness standards (e.g., (Madaio et al., 2020)); upon iterations, we addressed practical challenges that participants had explicitly mentioned or that I had implicitly observed during interviews. Our iterations allowed us to observe less obvious nuances in how practitioners operationalize fairness in later interviews in the study.

3.4.1 Designing the Prototype

Before conducting interviews, I created a low-fidelity prototype of the tools. The purpose of these prototypes was to help us uncover the practical challenges that practitioners might face when confronting fairness evaluation for the first time in a real recommendation setting.

To create these prototypes, I conducted a literature review to inform the creation of our decision tree prototype, particularly its high-level metric category leaf nodes. This literature review formed the foundation for designing a process needed to scope a quantitative fairness context and identify a fairness metric category for said context. I created the

original corpus by searching on the Google Scholar repository using the keywords "fairness," "ranking," "recommendation," "recsys," "fair," and "metric." I also added to our paper repository through snowball sampling; when I encountered a citation in one paper that referenced fairness metrics or definitions introduced in another paper, I added it to our repository. This process generated a corpus of 42 papers. I filtered this initial corpus to only include papers that introduced a new fairness metric. This removed papers that introduced re-ranking algorithms, or used previously defined fairness metrics. I also excluded papers that made assumptions that did not apply within practical large-scale recommendation contexts (e.g., if they assumed the practitioner had access to "ground truth" ranking labels) due to the motivation for this work to be useful in practical large-scale industry settings. This filtering process resulted in a total of 24 papers for the creation of our decision tree. Fourteen of these papers introduced provider fairness metrics, while nine of these papers introduced consumer fairness metrics. One paper introduced both provider and consumer metrics. It is important to note that this literature review should not serve as a comprehensive literature review since its goal was to scope metric *categories* for our prototype, not uncover every possible fairness metric. A list of these final papers that were leveraged for the creation of our prototype can be found in the "Literature Review" section after the References section of this dissertation.

Scoping harms for Interviews

Before interviewing participants, I asked them to complete an online questionnaire to begin thinking about which algorithmic harms they might want to explore during the interview. The questionnaire first asked participants to select a familiar recommendation or ranking system to evaluate. Next, participants were prompted to scope potential harms. I used these as the basis for navigating through the decision tree during their interview to scope quantifiable fairness objectives and identify their related metrics. Participants were

also asked to answer questions about their current role, their connection with the system they had chosen, and their comfort and knowledge of fairness metrics.

Prototyping the Decision Tree

Our original designs of the decision tree attempted to confront some of the challenges that previous research had uncovered regarding scoping quantitative fairness evaluations. For the decision tree, my collaborators and I wanted to help practitioners scope their predefined qualitative harm formulated from the questionnaire into a quantitative harm with respect to common fairness terminology. To do this, I designed various "decision-making nodes" with corresponding branches (decisions) to demonstrate the differences between constraints such as provider versus consumer fairness, individual versus group fairness, and ranking versus distributional fairness. These branches were designed as a way to organize the content I found during the literature review, where each branch corresponds with a related fairness constraint that I encountered in the literature.

The first decision-making node in the tree confronts the multi-stakeholder constraint, allowing practitioners to choose a branch between evaluating fairness for providers or consumers. The next decision node confronts the individual versus group constraint for both provider and consumer branches. For the providers branch, more decision-making nodes allow practitioners to select if they would like to measure fairness between multiple groups or for one group at a time. This design decision reflected current literature on provider group fairness, which introduces more nuanced measurements concerning groups.

The final decision-making nodes address system component (e.g., measuring fairness in item distributions versus in item rank positions) and fairness measurement property constraints found in our literature review. We did not designate a system component constraint branch for scoping consumer or individual provider fairness definitions due to the

lack of literature for those specific paths. The leaf nodes represent specific fairness objectives addressing the various constraints. For example, when navigating through the decision tree for a multi-group provider ranking fairness context, the final node might ask the practitioner to choose between two fairness objectives: (1) does the top-k ranking contain a specific proportion of items from these provider groups? or (2) are clicks, exposure, ranking, etc., proportional to item relevance for these provider groups? This choice of fairness objective leads the practitioner to a leaf node. Each leaf node has an associated "fairness category" that includes a list of metrics that fit within that category, both of which can be explored further in an associated spreadsheet. Tables 5.1 and **??** provide an overview of these fairness categories and their associated constraints, objectives, and metrics. The final design of the decision tree is the result of five iterations during our interviews, and one iteration once interviews were complete. I note that the decision tree is not meant as a final usable 'product', but rather as a starting point for more research and design. This version of the decision tree can be found in the Appendix in Section 7.4.

3.4.2 Interviews

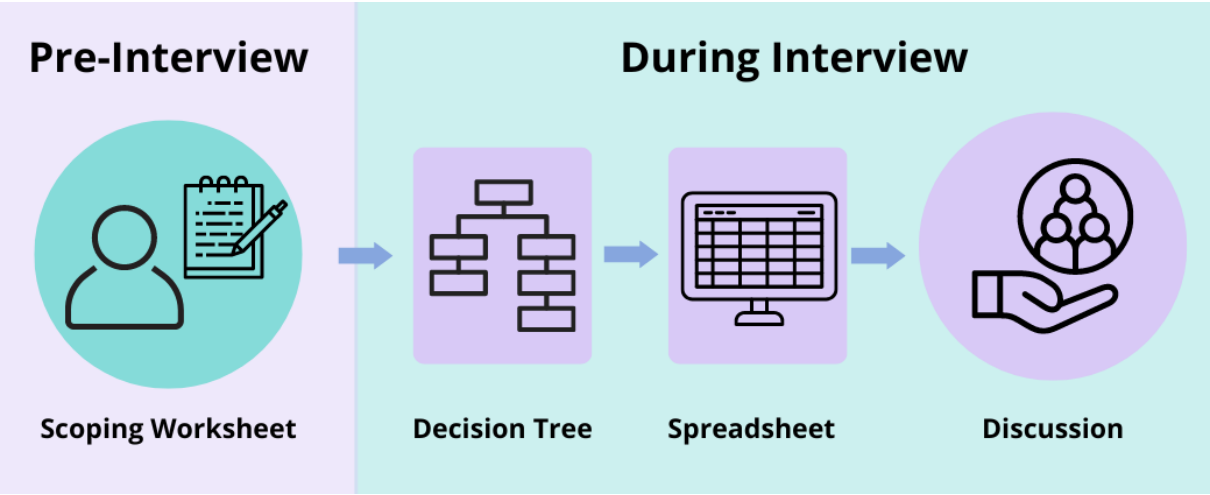


Figure 3.1: Interview process and artifacts used for Study #4.

I conducted a total of 15 semi-structured interviews with machine learning (ML) or ML-

adjacent practitioners working in content recommendation at Spotify. Each participant provided informed consent to participate in the study. All participants were working in a commercial consumer-provider recommender system setting with millions of users worldwide. The participants referenced in this paper can be found in Table 3.3, while the flow of the interviews and the associated artifacts used can be found in Figure 3.1.

Table 3.3: Participants from Study #4 and their roles. The "Participant ID" column corresponds to the alias that I use to identify each participant in Section 5.3.

Work Area	**Roles**	**Participant ID**
Machine Learning	ML Engineer, ML Engineer Manager, Product Manager	P1, P2, P5, P6, P7, P8, P3, P13
Data Science	Data Scientist, Data Science Manager	P4, P9, P10, P11
Research	Research Scientist, UX Researcher	P12, P14

All interviews were conducted online via Google Meet, and participants were not additionally compensated for their voluntary participation. Rutakumwa et al. (2020) showcased that opting to not record interviews while maintaining robust written documentation can preserve participant privacy while ensuring results do not change. As such, our interviews were not recorded to allow for participant privacy; all quotes from participants were gathered from notes taken during interviews. Participants were recruited via snowball sampling, where previous participants connected the research team with other relevant practitioners who they had worked with who might be a good fit for the study. We conducted inductive, thematic analysis (Clarke et al., 2015) on interview notes, and categorized participant responses and observations into themes and sub-themes.

3.5 Study #5: The Experts' Perspective

Collaborator Statement: For this study, I had the privilege of collaborating with Michael Madaio, Robin Burke, and Casey Fiesler. My contributions to the research included

designing and conducting all interviews, leading the analysis and coding for all interview transcripts, and leading the writing for the resulting paper.

I conducted semi-structured interviews with 21 ML fairness experts from industry and academia to understand: (1) how they currently conduct ML fairness evaluations; (2) how they would improve their current processes for conducting fairness evaluations; and (3) what resources and mechanisms they think would be necessary to make those improvements. The interview protocol for this study can be found in the Appendix in Section 7.4. Interviews were 60 minutes, and all participants were given the option to be compensated with a $30 digital gift card for their participation. This study was approved by the University of Colorado's Institutional Review Board.

3.5.1 Participants

I recruited 21 participants from industry and academia, who self-identified as *"ML fairness experts,"* and/or *"recommender system experts."* The inclusion criteria for participants from industry was they must have participated in Responsible AI (RAI) or ML fairness efforts of some kind; in academia, they must have published papers about ML fairness or RAI at conferences such as FAccT, CHI, CSCW, or other similar venues. I included academics in this study to ensure our results were not limited to perspectives of those currently working within industry, and to more easily recruit participapnts with interdisciplinary ML fairness expertise. Our inclusion criteria also included people who have technical experience with recommender systems, either in academia or in industry; for these participants, no prior knowledge of fairness or RAI was necessary.

I recruited participants first by convenience sampling: sending emails to people who I personally knew had conducted recommender system fairness research and/or had done ML fairness research in industry or academia. I also performed a version of snowball

sampling with participants by asking them to share the recruitment call with others who they thought would be a good fit. I conducted three pilots of the interview study, and then conducted 21 official interviews with recruited participants. All interview participants filled out a pre-interview survey where they described their previous experience with ML fairness and recommender systems, and provided their demographic information, shown in Table 3.4. I recruited 10 participants from academia, and 11 participants from industry, with varying years of experience with ML fairness, RAI, and Recommender Systems topics. I report the age, gender, and race/ethnicity of our participants in aggregate in order to avoid identifying individuals. The median and mean age of our participants was 37, the minimum age was 27 and maximum age was 58 (3 did not answer); the self-reported gender distribution of our participants was 7 Females, 10 Males, and 1 Nonbinary person (3 did not answer); and the self-reported race/ethnicity of our participants included: White, Middle Eastern, Filipino, Austrian, Jewish, Latino, West Asian, Asian, and Hispanic (4 did not answer). Some participants reported being a combination of several of these races/ethnicities, I decoupled these combinations in order to avoid identifying individuals.

Table 3.4: Expertise of participants in Study #5. ML=Machine Learning, RAI=Responsible AI, RecSys=Recommender Systems.

P#	Affiliation	Self-Described Expertise	Years of Experience With...		
			ML Fairness	RAI	RecSys
P1	Academic	Computer Science	7	2	7
P2	Industry	HCI	7	20	20
P3	Industry	Data Science, HCI	5	5	0
P4	Academic	Computer Science	10+	10+	29
P5	Academic	HCI	8	20	20
P6	Industry	Computer Science	6	6	6
P7	Industry	Computer Science	8	NA	10
P8	Academic	HCI, Computer Science	12	12	17
P9	Industry	Data Science	6	7	3
P10	Industry	Data Science Management	6	6	10
P11	Industry	Computer Science	NA	NA	Some
P12	Academic	Computer Science	5+	5+	10+
P13	Industry	Computer Science, Mathematics	7	7	9
P14	Academic	STS, Social Science	5	5-10	5
P15	Industry	Computer Science	5	5	2
P16	Industry	HCI, Computer Science	4	5	6
P17	Industry	Data Science, Applied Research	4	4	0
P18	Academic	HCI, Computer Science	9	9	15
P19	Academic	Qualitative Social Science	8	8	0
P20	Academic	Anthropology	10	NA	15
P21	Academic	HCI	7	10	1

3.5.2 Data Analysis

After completing the interviews, I conducted reflexive thematic analysis, starting with inductive open coding of all transcripts, which yielded 62 initial codes (e.g., "benefits of user studies" or "choosing a threshold"). My collaborators and I then grouped these codes into themes by clustering them into topic areas (e.g., "methods for conducting fairness evaluations" or "incorporating users into fairness evaluations"), discussing them with colleagues to identify the most salient themes. Our analysis followed reflexive thematic principles, treating data themes as stories about shared meanings and data topics as domain summaries without shared meaning-making (Braun and Clarke, 2021a). Afterward, I conducted Member Checking to verify quotes and demographic details (Birt et al., 2016).

3.6 Methodological Limitations

Although the methods used in this dissertation allowed me and my collaborators to explore the perspectives of various stakeholders of recommender systems, the qualitative nature of this research has various limitations.

First, while semi-structured interviews and focus groups allowed us to learn about the lived experiences and needs of stakeholders, it also does not allow for generalizability. This means the resulst of this dissertation can provide key insights into potential best practices based on the needs of our study participants, but do not act as a holistic guide that covers all users and practitioners needs for fairness in recommender systems. It has been shown that the average sample size for qualitative studies in HCI is 12 (Caine, 2016), and the sample sizes for the five qualitative studies included in this dissertation was: 30, 13, 23, 15, and 21, respectively. Although the sample sizes of these studies were relatively small, they still allowed for us to elicit enough information to learn shared meanings about

fairness operationalization from the participants.

Golafshani (2003) describes how the concepts of validity, reliability, and triangulation align with the principles of qualitative research, which can eventually lead towards notions of generalizability when applied well in practice. However, this author notes that the focus of qualitative research is not on generalizability, but rather about illuminating and understanding shared meanings for similar situations. In this light, the limited number of participants in these studies might not represent the broader stakeholder and demographic populations they are a part of. In addition, I do not claim that data, themes, codes, or the meanings I derived from them were "saturated" in our studies. Braun and Clarke (2021b) describe how in thematic analysis, saturation of themes and meanings is not the goal, nor is it possible to derive a generalizable point during the qualitative research process where saturation has been achieved. In grounded theory, data saturation is a goal because it indicates that no new themes, categories, or insights are emerging from the data, signaling that the researcher has thoroughly explored the phenomenon under investigation. However, in thematic analysis, saturation is not the goal because the approach is not inherently tied to generating a fully exhaustive set of themes but rather focuses on identifying patterns of meaning within the data that align with the research questions and the analyst's interpretive framework (Braun and Clarke, 2021b). Further, reflexive thematic analysis (as is the primary method I adopted for this dissertation), is considered a reflection of the researchers' interpretations of the data, and it is expected that codes and themes produced by one researcher may not be replicated by another (Byrne, 2022).

Despite these limitations, the perspectives that were derived from participants are still valuable and remain deserving of exploration because they provide nuanced, context-specific insights into the lived experiences and challenges faced by individuals directly impacted by the phenomena under study. These insights can illuminate unique perspectives that may otherwise be overlooked in broader quantitative analyses, offering meaningful

contributions to the understanding and practice of fairness in machine learning.

With all this in mind, there are additional limitations to the insights produced from the qualitative research in this dissertation. First, response bias (Furnham, 1986) may have influenced participants' answers across the semi-structured interviews and focus group study, as individuals might have been inclined to provide responses they believed were more socially acceptable or aligned with the expectations of the interviewer or other participants, rather than offering their true feelings or experiences. This was likely compounded by social desirability bias (Bergen and Labonté, 2020), where participants could have framed their responses in ways that they believed were socially favorable or appropriate for the context. I expect this bias to have been especially present during the focus group study, due to the fact that participants were in a zoom room with one another, and after introducing themselves to each other, might have felt uncomfortable to disagree with people they did not know well. However, I will note that in one of the focus groups, two participants did debate one another about the ethics of allowing AI-generated art on social media platforms, which showcased that for at least some of the participants in that study, they felt comfortable naming when they disagreed with the values of another participant.

Additionally, a common challenge in qualitative research, particularly in the interviews and focus group, is the difference between theoretical preferences and practical realities. What participants expressed in theory about fairness, equity, or other concerns might not have always aligned with what they actually prioritize or value in practice, even if they were unaware of this gap. Related to this, participants might not always have had a full or accurate understanding of their own needs and priorities. As a result, the insights gathered from these studies, particularly when participants described what they wanted or needed, may not always fully capture their true preferences or the most effective solutions for their contexts.

Finally, I acknowledge that despite conducting five studies for this dissertation, I was not able to address every research question I had hoped to explore. While I discovered that practitioners need more guidance when scoping fairness evaluations and determining which stakeholders and fairness goals to prioritize, I have not yet identified best practices for providing this prioritization guidance. Additionally, I found that designing fairness metrics in collaboration with impacted stakeholders shows promise as a method for improving fairness design, but I have yet to implement this approach in practice. Such an implementation is likely to introduce new challenges that remain to be fully understood. While I have examined some best practices for conducting pragmatic fairness in industry settings, these insights are limited to the 21 participants of my final study. Expanding this research to include more participants from diverse organizations and areas of expertise would likely yield a broader array of approaches to advancing pragmatic fairness efforts. Finally, the scope of this dissertation is necessarily constrained by its focus on recommender systems. This allowed for an in-depth exploration of fairness operationalization within a specific context, yet many other domains of machine learning could benefit from similar research. I encourage future studies to apply the methods and questions explored in this dissertation to these other domains to expand our understanding of fairness in machine learning more broadly.

3.6.1 Positionality Statement

My identity as a White, queer, middle-class, American woman with a degree in software engineering, living in Boulder, Colorado, has shaped the research presented in this dissertation in many ways. Throughout my research, I was conscious of how my personal background, experiences, and perspectives influenced the choices I made in all stages of the research process. The research questions I asked, the design of the interviews and focus groups, the types of follow-up questions I chose to ask during studies, and the codes

and themes I identified as salient were all influenced by my lived experiences, personal interests, and inherent biases.

As someone with a technical background in software engineering, my perspective on fairness in recommender systems was shaped by both my familiarity with the limitations of technology and my awareness of the importance of human-centered design in the tech industry. My academic training and industry internships have also made me sensitive to how fairness frameworks are operationalized in practice and the challenges practitioners face when trying to apply theoretical concepts to real-world systems. Living in Boulder, a progressive community with a strong academic and tech presence, may have further influenced my thinking, as it also exposed me to specific perspectives on fairness, ethics, and technology.

Additionally, my identity as a queer woman influenced the way I engage with stakeholders and impacted my commitment to promoting fairness for underrepresented or marginalized groups in technology. As a result, I intentionally centered my studies on understanding the needs of diverse stakeholders, aiming to highlight the voices and perspectives that are often overlooked in fairness discussions. As a middle-class white woman, my experiences have been shaped by privileges that may have shielded me from the harms and limitations faced by individuals with different identities. I remain conscious of these privileges and the ways in which they may have influenced my interactions with participants and the interpretation of the results in my studies. This awareness informs how I approach the findings, recognizing that my perspective may not fully capture the lived experiences of those who face different societal challenges.

Overall, I acknowledge that my positionality has shaped both the design and analysis of my studies, influencing the way I interpreted data and framed the significance of the findings. I remain mindful of these influences and hope that my reflections on them can help contextualize the insights and recommendations that emerge from this dissertation.

Chapter 4

The Users' Perspective

Overarching Research Question: How do users think that fairness should be operationalized in real-world recommender systems?

In this chapter, I explore how we can better operationalize fairness *with* users of recommender systems. I discuss how users reason through fairness considerations, how they may be impacted by them, and how they wish that practitioners operationalized fairness for recommender systems. I present my findings from two studies: (1) an interview study with end-users of recommender systems; and (2) a focus group study with providers of recommender systems. I first begin by providing some theoretical background and motivation for these studies, and then share the results of each study, as well as the implications of these results.

4.1 Motivation

Previous work exploring best practices for machine learning fairness has largely focused on the perspectives of the practitioners (e.g., (Madaio et al., 2020; Holstein et al., 2019;

Rakova et al., 2021)). While these perspectives are useful and necessary to aid in the improvement of fairness operationalization–they omit several key stakeholders from the design process: those who are *impacted* by fairness in the system. Costanza-Chock (2020) describes that one of the core tenets of community-based design is to build technologies that might impact communities *with* those communities who may be impacted.

Fortunately, in an effort to better align recommendation algorithms with users' needs, researchers have begun to turn towards participatory and collaborative methods (Delgado et al., 2023; DeVito et al., 2021). Many previous studies have explored fairness perceptions in AI; one study explored 200 papers that had focused on this effort—however the authors of this work found that many of these prior studies were limited in size and scope (Van Berkel et al., 2023). A similar study that explored 58 prior studies on people's perceptions of algorithmic fairness found similar limitations of prior work, particularly with respect to viewing algorithmic fairness through a western-centric lens (Starke et al., 2022).

Jannach and Bauer (2020) describe how conducting research with users is important to validate the success of a recommender system, rather than optimizing the system for metrics without understanding how those operationalizations might impact real people. This research also describes how the original intention of recommender systems was to make life easier for consumers of platforms, but that this excludes several other stakeholders in the system, such as item providers. They propose that recommender systems ought to create value to not just the consumers of recommendations, but also the providers. In addition, Stray et al. (2022) describe how designing metrics with system stakeholders can help make systems more robust, and better aligned with human values.

Konstan and Terveen (2021) describe the goal of **human-centered recommender systems** research as the following: *"to design the algorithms and interactions of recommender systems to better fulfill the goals of users and of the organizations engaging with these users."*

These authors describe a scenario of a very successful recommender system for grocery stores that learns to recommend bananas and bread to everyone. On the surface, this recommender is functioning very well—it reports high accuracy and almost always recommends items that a customer buys. Underneath the surface, it is obvious that this recommender provides little value to the customers of a grocery store. Relying solely on metrics (accuracy or otherwise) tells us very little about the real experience that stakeholders of a recommender system are having. In the case of fairness, aligning a recommender systems' objectives with those of the stakeholders' is a higher-stakes process—to do so poorly could result in much dire consequences than poor user experience. Previous work has also described how overemphasizing metrics in the evaluation process has become a fundamental problem in the discipline of machine learning (Thomas and Uminsky, 2020).

Birhane et al. (2022) describe that conceptual analysis of a system can only take us so far as ML researchers. To really target and improve structural and historical power asymmetries, research in the discipline of AI ethics ought to focus more on the real experiences of real people. One approach to confront this challenge is to involve users in the design of ML systems and their objectives, such as fairness.

Recent research within the computer science discipline has shown the positive impact of including users in the design process of fair machine learning systems. Srivastava et al. (2019) describe that all stakeholders of fairness-aware algorithmic decision-making systems should be involved in the fairness formulation process.

> "Algorithmic decisions will ultimately impact human subjects' lives, and it is, therefore, critical to involve them in the process of choosing the right notion of fairness... After all, a theory of algorithmic fairness can only have a meaningful positive impact on society if it reflects people's sense of justice," (Srivastava et al., 2019).

Lee et al. (2019b) showed that involving users in the design of an algorithm led them to perceive and believe the resulting algorithm to be fair, and that this co-design process improved user attitudes towards the platform as a whole. The same results have been shown in research on procedural justice (Lind and Tyler, 1988; Lee et al., 2019a) and participatory policymaking (Fung, 2003).

As I previously noted, in the domain of recommender systems, end-users are not the only stakeholders who are impacted by fairness. *End-users* or *consumers* are the stakeholders who consume recommended items, and their fairness interests and goals might differ substantially from *Providers*—those who *provide* items to be recommended. In this chapter, I involve both of these stakeholders in the design process of fairness operationalization, as both of these stakeholders are "users" of the system in their own right.

4.2 Study #1: The End-Users' Perspective

Fairness and Transparency in Recommendation: The Users' Perspective

(Sonboli, Nasim and Smith, Jessie J. et al., 2021)

4.2.1 Introduction

In this preliminary work, my co-authors and I explored what end-users' perceptions of fairness are when it comes to recommendation, and how they would like to be informed about fairness interventions that might impact their personalized recommendations (Sonboli, Nasim and Smith, Jessie J. et al., 2021). We also explored the different opinions that end-users of recommendations may have when it comes to their personal definitions of fair treatment on recommendation platforms (Smith et al., 2020). This was a semi-structured interview study with 30 participants who had interacted with recommender systems before. More details about the methodology of this study can be found in Section 3.1.

The research questions for this study were the following:

- **RQ1:** What are end-users' folk theories about recommender systems?
- **RQ2:** How do end-users of recommender systems want fairness goals and interventions to be explained to them on the platform?
- **RQ3:** What explanation designs do end-users like best for explaining how fairness interventions impact their recommendations?

4.2.2 Results

The results from this study highlighted that end-users would like to be informed about how fairness is incorporated into the systems they interact with, and that transparency around fairness considerations can be a useful tool to improve trust between end-users and recommendation platforms.

In this section, I describe the main findings from the exploratory interviews with users of recommender systems. I begin by explaining some of the gaps in participants' knowledge about recommender systems and fairness objectives, as gathered from participant folk theories. Then, I explore the ways in which participants indicated that explanation design of a fairness-aware system could help or harm their understanding and trust of the system. Participant quotes are indicated by anonymized participant numbers; participants 1-15 were studying CS or adjacent fields, and 16-30 were in non-technical fields.

Folk Theories of Recommender Systems

Some participants expressed that the "black box" nature of recommendations made it difficult for them to learn how a recommendation system might create personalized recommendations for them. For example, as P5 said, *"recommender systems in most cases are pretty unknown to the user."*

However, in general, participants had somewhat of an understanding for how recommender systems worked–particularly in terms of how much data was being collected on them. Some participants' folk theories aligned well with accuracy-based recommendation algorithms, such as collaborative filtering algorithms like User-KNN or Item-KNN.

> *"So if like a lot of people buy, you know, a hammock, then if a large subset*

of those people who bought a hammock also bought a sleeping bag, it might recommend the sleeping bag to you" (P21).

Folk Theories of Fairness Objectives

When asked about how recommendations could be unfair to users, very few participants indicated that they had ever thought of provider fairness for recommendations. This raised a concern that there might be a lack of communication between the system and the user when it comes to issues of provider fairness in recommendation.

After a short discussion about fairness objectives and the impact that a recommender system might have on the providers, many participants indicated that they thought provider fairness was important to them in a recommender system, as expressed by P18.

"I don't think I've thought about it as much from a seller's perspective, but I can see like, you know, these platforms are made for more than just [users]. They're made for [providers] as well. So like people who sell things on Amazon or make music and put it on Spotify, it's like, yeah, you're probably inherently at a disadvantage to people who are already big or are already [favored by] the algorithm" (P18).

However, the majority of participants were still not entirely sure how fairness goals might impact their recommendations. Thus, we began to explore different ways that a recommendation platform could explain this impact to users in an educational and empowering way.

Explaining Fairness Goals to Users

Throughout the interviews, most participants indicated that they would want to see the fairness goals of an organization described to the users in some way, either through a short explanation or an entire page, as described by P8.

> *"[Organizations] should have [fairness goals] somewhere I could find it like the little 'about us', like 'learn more about our corporation' tab where they would explain 'these are our moral values, these are what we prioritize'"* (P8).

One participant thought that fairness goals were best incorporated into the UX of the platform, as long as the language did not unintentionally manipulate users.

> *"Or like work [fairness goals] into their platform somehow. Like how Spotify does the 'New Music Fridays'... I think that there are opportunities to weave that into places on a platform but not necessarily blanket it across so the user doesn't feel like they have choice"* (P29).

This raised an important concern that is not new in the field of explanation: the concern that explanations, if not designed carefully, could be used as a tool to manipulate users.

Fairness as a User Choice

One outcome of fairness-aware recommendation is to "nudge" users into choosing items that they might otherwise not. The intention of this kind of system is to encourage the user to choose the items that meet the fairness objective of the system itself, which could unintentionally lead to "manipulation," as some participants noted. In fairness-aware recommender systems, *coercing* users into choosing a "fair" recommendation – which might

not be 'fair' for everyone – could be misleading. This concern was expressed amongst the participants, particularly because fairness definitions are often disputed, and what might be considered *fair* for some, could seem *unfair* for others (Smith et al., 2020). In order to remedy these concerns, several of the participants indicated that they would prefer to be informed about the existence of fairness-aware recommendations so that the user could make the choice of fairness or personalization for themselves, as expressed by P22.

> *"[Fairness-aware recommendation] is manipulative in some sort of way. I think the best thing that they can do would be to give a short explanation that they changed [the personalized recommendation algorithm] and then kind of show the whole list and not point out any others in the list... allowing an unbiased choice from the viewer"* (P22).

If manipulation was a concern amongst participants, then how should explanations be designed to mitigate that concern? Further, how might explanations instead be designed to foster trust and communication between a system and its users? In the next section, I propose that these concerns can be alleviated through explanations if they are used effectively as a tool for education.

Explanations as Education

One prominent theme that emerged throughout the interviews was that *transparency as a means for education was essential*. Whether an algorithm uses fairness objectives or not, participants expressed that they needed to be educated about why they were being recommended the items that they were. This point was succinctly summed up by P21 who stated: *"the more transparency, the better."*

Many participants indicated that better design practices could help promote fair treat-

ment for providers and relay this fair treatment back to the recommendation consumers. In order to ground participants in a specific platform and discuss concrete explanation designs, we used Kiva as a case study. More information about the Kiva organization can be found in Section 3.1.1. Specifically, we asked the participants how they thought transparency could be designed when fairness goals were incorporated in a system like Kiva. We gathered feedback about the benefits and flaws of explanation designs that were minimal and global, versus detailed and specific.

On Kiva, recommended "items" are actually people – borrowers looking to get microloans funded for necessities such as food, water, and medicine. When a borrower is closer to the top of the recommendation list, they are more likely to have their request clicked on and potentially funded by lenders (Choo et al., 2014). Fairness-aware recommendations on Kiva shift where a borrower lands in a consumer's recommendation list – sometimes raising their position in the list, other times lowering their position in the list – which presents high stakes for the providers on this kind of platform.

During our interviews, we asked participants to pretend that they were a lender (and thus a consumer of recommendations) on Kiva. We presented a scenario where fairness-aware recommendations changed which borrowers appeared in their personalized recommendation list and asked them how they would want this change to be translated to them as a consumer. As described in the methods, we presented three scenarios for explanation design.

The first option, illustrated in Figure 4.1 (a) was to provide no explanation of any differences in their recommendations due to fairness objectives. The second option, illustrated in Figure 4.2 (b) was to provide a general fairness explanation at the top of the recommendations to inform users that some of their recommendations could be different due to ambiguous "fairness goals". The third option, illustrated in Figure 4.3 (c) was to provide specific explanations for each recommendation that was ranked higher in the list

due to fairness goals, and why that recommendation met Kiva's fairness goals. These scenarios allowed us to explore consumer preferences in terms of the number and depth of explanations in fairness-aware recommendations. Finally, we asked users if they would rather explanations be kept out of their recommendations and instead have the option to simply select between purely personalized recommendations and "fair" recommendations, through the use of a toggle button.

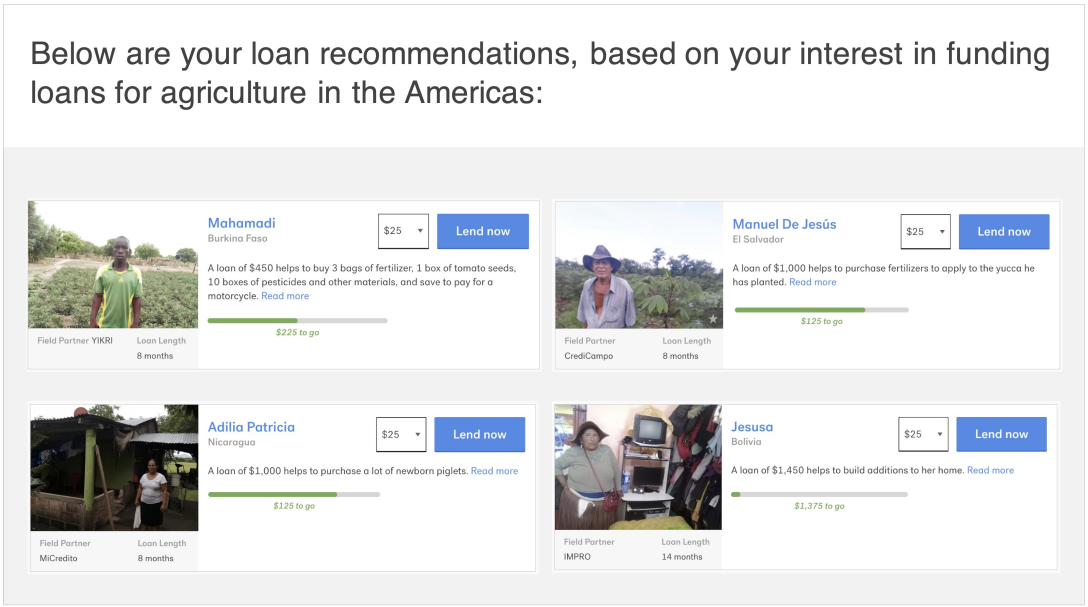


Figure 4.1: The various explanation designs we presented to participants during our study.

Transparency Without Explanations

Most participants indicated that if the recommendation algorithm on the Kiva platform (or other similar fairness-aware platforms) were to change due to fairness goals, they would need to be notified. Thus, participants generally expressed that transparency without explanations was not ideal.

> *"I would probably expect them to tell me the values that were kept in mind when coding [the fairness-aware recommendation algorithm]... I wouldn't expect to*

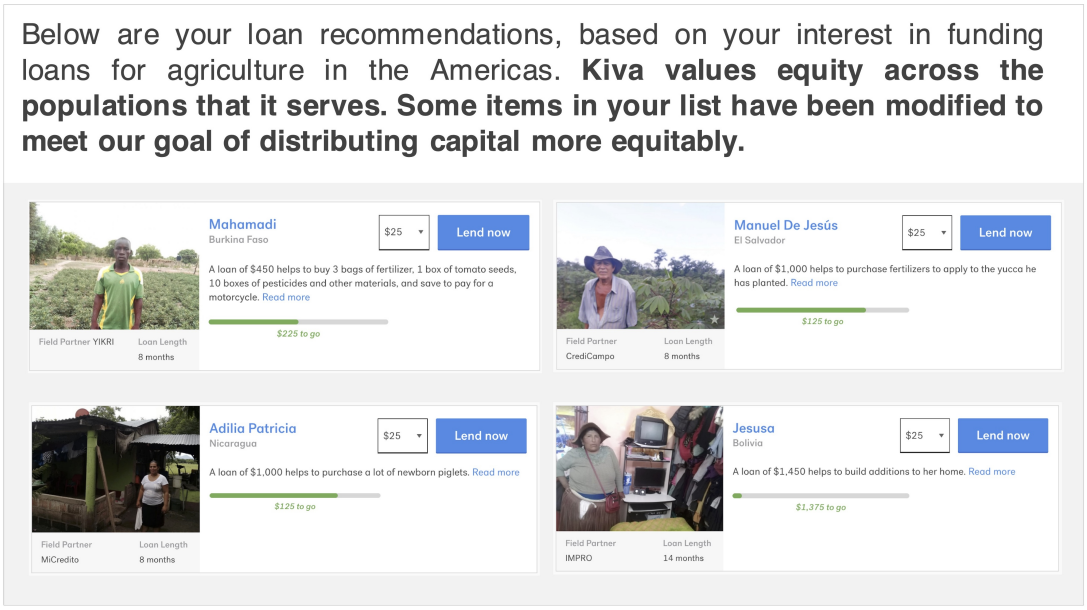


Figure 4.2: The various explanation designs we presented to participants during our study.

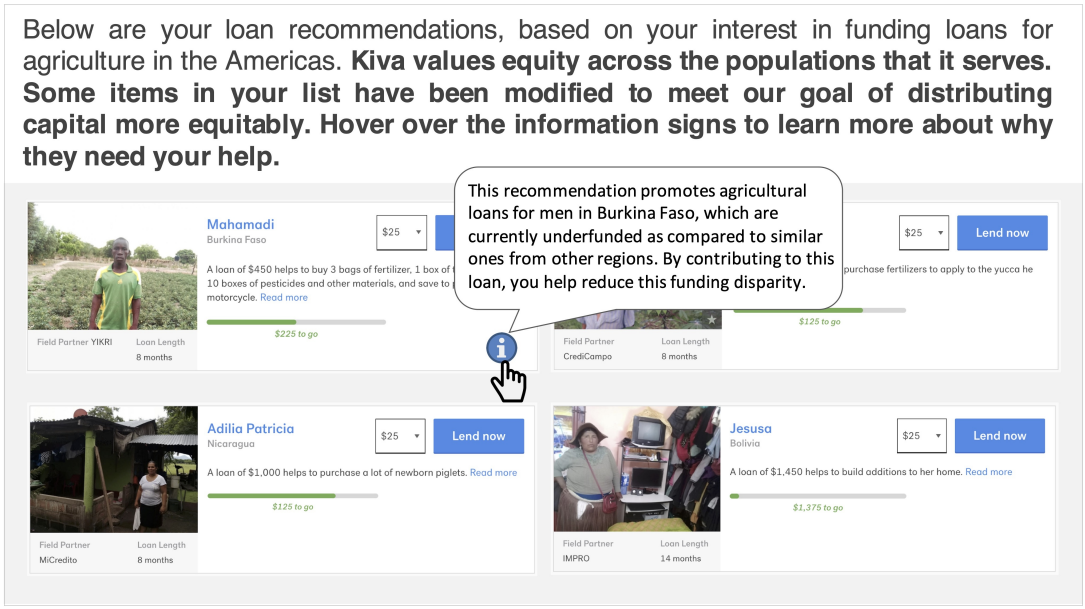


Figure 4.3: The various explanation designs we presented to participants during our study.

hear nothing that's for sure. I expect to hear something" (P8).

For many of the participants, explanations were expected in the design of a fairness-aware system.

Transparency With Minimal Explanations

Given that all participants indicated that they needed *some* explanations of fairness goals, the remaining options for explanation were either one generalized explanation, many detailed explanations, or somewhere in between.

> *"If [fairness-aware recommendation] is incorporated, I don't think I would need them to explain every time, but maybe like at the beginning, just understand they have a [fairness] goal and then go on from there as normal"* (P8).

Several participants were in favor of including one global explanation for all fairness-aware recommendations. Some indicated that a single explanation could still leave room for the system to include more in-depth explanations about a specific group of providers that the system is optimizing for fairness. For example, many participants expressed that they would like to know more about why Kiva borrowers from a specific geographic area are underfunded on the platform, and how choosing to lend to people from that area could have a positive social impact.

> *"I like [longer explanations], where you can scroll underneath and get like a broader perspective of what's going on in the whole region, I would like that"* (P5).

Transparency With Detailed Explanations

Another alternative was to provide a separate explanation for every individual recommendation, on the recommended item itself. Many participants expressed that if explanations were specific to each recommendation, the manner in which they were formatted played a big role in their utility to the consumer. P30 described that for explanations to be effective and informative, the *"algorithms should be explained to you in a simple sentence structure... short, concise and in digestible pieces"*. They added that explanations should be accessible and inclusive for all audiences, regardless of their level of understanding of algorithms.

In summary, my collaborators and I found that when it comes to transparency for fairness-aware recommendations, design decisions matter. Transparency should give users greater agency to understand how provider fairness might make their recommendations different than personalization. Further, when recommendation fairness is explained to the consumer, it should be explained in a way that is accessible, understandable, specific, and concise.

4.2.3 Discussion

Taking into account the feedback provided by the interview participants in Study #1, the following are suggestions for features that could be included in effective explanation design for fairness-aware systems.

1. ***Explanations should define the system's fairness objective for users.*** For example, an explanation could educate the user about the impact that a fairness-aware recommendation might have on the item providers.

2. ***Explanations should not nudge/manipulate users into making a decision, even if the goal is fairness.*** For example, an explanation could educate users about the existence of a fairness-aware recommendation while still allowing the user to decide if they prefer personalization instead.

3. ***Explanations should disclose the motivation for using fairness as a system objective.*** For example, an explanation could educate users about the fairness concerns of the system, or the values of the organization and why they chose fairness as an objective.

These three features were compiled based on the specific gaps that the participants had in their knowledge about fairness-aware recommender systems. Specifically, the folk theories that participants shared on these systems led us to believe that there might exist a major gap in communication towards users about recommender systems' objectives. These features also summarize participants' most frequent desires for what they wished explanations could teach them about the system in order for them to understand and trust it more. It is also interesting to note that whenever participants were educated about provider fairness concerns in recommendations, they generally showed interest in the option of using a system that had fairness as an objective.

4.3 Study #2: The Providers' Perspective

Recommend Me? Designing Fairness Metrics with Providers

(Smith et al., 2024)

4.3.1 Introduction

In this work, I explored the lived experiences of fairness from the perspective of **providers** (people who provide items for recommendation), and co-designed metrics that empirically captured these experiences of fairness. This research provides a concrete method to assist ML practitioners with developing fairness definitions and metrics that are aligned with their users' needs, which previous work in the FAccT community has explicitly called for (Deng et al., 2023). During four focus groups with thirteen total participants, I co-designed fairness metrics with providers from two domains of recommendation: content creators and dating app users. More details about the methods of this study are described in Section 3.2.

The research questions for this study were as follows:

- **RQ1:** What are the lived experiences and perceptions of unfairness for content creators and dating app users?
- **RQ2:** What fairness goals, definitions, and metrics do these providers want enacted in their recommender systems to alleviate these experiences of unfairness?
- **RQ3:** What are some of the opportunities and challenges of designing fairness metrics with providers from different recommendation domains and contexts?

4.3.2 Results

During the focus groups, I first discovered that most participants were uncertain about how recommendation systems worked. This lack of understanding also influenced participants' abilities to recount their experiences of algorithmic fairness. For example, PD5 said *"I think it's kind of hard to say whether [I'm] being treated fair or unfairly just because I don't really have a good understanding of how the algorithm works overall."* PD4 also shared that they felt *"uncertain about the level of fairness,"* that they experience, *"just because of the opacity of the like algorithm itself."* Although uncertainty about the algorithm's functionality made it difficult at times to understand if the recommender system was being fair, participants still had many experiences of unfairness to share and discuss with one another.

In the following sections, I describe these experiences and how they informed the participants' designs of their fairness goals, definitions, and metrics. Though there were commonalities between the two domains, overall there were many contextual differences, so I categorized these results by the participant type (Content Creators or Dating App Users) and noted when any similarities were observed between both groups.

Lived Experiences of Unfairness

When asking participants about their experiences of fairness and unfairness on their associated platforms, several participants struggled with interpreting the word "fair," such as PC1, who expressed that it's *"hard to say what is unfair, just what **feels** unfair"* (PC1). As noted in Section 3.2.2, I did not provide a definition of fairness for participants, nor did I explicitly elicit one from participants. Although this decision prevented me from ensuring that responses aligned with a mutually agreed understanding of what "fairness"

is, this absence of a definition also allowed us to explore what the participants considered to be "unfair" based solely on their personal experiences and preferences.

Content Creators

For content creators, one experience of unfairness that emerged was inequality of content virality or exposure. PC1 shared that they felt the algorithm prioritized creators with a large following, regardless of their quality of content: *"I think you know the sort of narrative that... if you [make] good content that people want to see, you will go viral... however I see many of times that big creators do the same as small creators and go viral for it"* (PC1).

PC5 expressed discouragement that their profile might never be fairly exposed by the algorithm, given their small follower count: *"when I look at the profiles of other people who are doing the same thing, they frequently have 10's of thousands of followers. I don't feel like I'm ever going to get there when I only get one or two new followers per day"* (PC5).

Another experience of unfairness recounted by the participants was related to algorithmic content suppression and shadowbanning (algorithmically blocking/suppressing provider's content without their knowledge) (Gillespie, 2022). *"My observation of Instagram is that it's suppressing everybody all the time. When they switched to reels it was clear that they had no interest in pushing image content at all"* (PC3). PC4 and PC6 both shared experiences where their content was removed for being *"too sexual,"* even though it was them *"dancing fully clothed"* (PC4). These participants further discussed how their content had been banned (PC6) or algorithmically suppressed (PC4) without explanation. PC6 noted one instance where a post was removed *"about [not] blaming victims of sexual assault and telling them that they need to be aware that they could get assaulted on the subway,"*

and that they were not sure why this content was banned.

When these participants were asked *why* these experiences felt unfair, they described that it led to feelings of frustration (P3) and upset (P6), or caused them to lose opportunities when their content was incorrectly suppressed without any warning or rectification (P4). In summary, for the content creators, unfairness was mostly experienced when their content did not receive enough exposure on the platform, or when their content was suppressed by the recommendation algorithm—without any understanding as to why this was happening, nor the agency to prevent it.

Dating App Users

For the dating app participants, one shared experience of unfairness was related to unwanted profile exposure that did not align with their stated preferences. For example, PD1 said that *"I want to see [women] in my dating apps, and somehow men keep seeing my profile and keep liking my profile in a way where I genuinely don't understand, like how that happens"* (PD1). This participant said that this mismatch between preferences and exposure felt unfair because it was wasting both their time and their prospective matches' time, and possibly leading to a loss of opportunity. *"If there's people who are like a categorical rejection for me, like, I will categorically reject men on dating apps, then it's like wasting my time. It's wasting their time... it kind of feels like something has been sort of like taken away, or like you've lost an opportunity"* (PD1).

Similarly, PD3 shared that even though they turned off preferences for men on their Hinge app, they still received likes from men and wondered *"how the hell did you even see me? Because I'm, like, turning [that preference] off. Like I don't want to talk to men"* (PD3). This participant said the inability to expose their content based on their explicitly stated preferences felt unfair to them because it seemed like the algorithm was making

assumptions about their identity and sexual preferences based on their appearance, which felt binarist.

> *It feels unfair to me. Like I present extremely feminine, and I know that... But like, I feel like it's taking the way I look, the way I present for my picture... and then completely filtering out an entire like group of people by saying 'well like even if someone is nonbinary, if they look a certain way, we're going to put them in the same category as [men or women]'. It just feels like it's binarist without saying it is, even if it gives you that third gender option* (PD3).

PD2 added an additional consequence of this fairness concern; when their profile had been exposed to people outside of their stated preferences, this led to hateful speech in their direct messages. *"I had one time this very racist person send... something like very racist... So that was weird, like if my preference is not [this person] why are you sending me likes from these people?* (PD2). PD5 similarly described that when their profile was exposed to people who did not align with their preferences, they would *"get ignored or kind of talked down to,"* and this led them to feel like the platform was *"not really a safe space"* (PD5).

In summary, all of these experiences showed that when dating profiles are algorithmically exposed to an incorrect audience, this can lead to feelings of discomfort, loss of opportunity, feelings of discrimination or exclusion, and concerns for safety on the platform. Interestingly, it appeared that many fairness concerns for content creators were about under-exposure, while many fairness concerns for dating app users were about over-exposure—implying that fairness goals, definitions, and metrics designed to capture these experiences might need to measure different effects.

Fairness Goals and Definitions

After discussing the participants' experiences of unfairness on their associated platforms, I asked them to complete a series of activities to develop fairness goals and definitions that might improve their experiences of fairness on a given platform. Examples of some of the fairness goals and definitions that participants developed are shown in Table 4.1. I discovered two categories of fairness goals and definitions that were shared between dating app users and content creators: (1) exposure equality; and (2) transparency.

Content Creators

- ***Exposure Equality.*** Several participants designed fairness goals and definitions to improve equality of content exposure. For example, PC4's fairness definition was to provide *"equal opportunity for all [creators]"*, while one of PC3's fairness goals was to improve *"exposure consistency"* (PC4). PC1 similarly defined fairness as not *"demonetiz[ing] shops or business accounts, [don't] reward people with high views or daily content [and] stop suppressing creators that aren't white men, women who show skin, or non-cis creators."* PC1 and PC2 (who had discussed that it felt unfair when their followers weren't being recommended their content) designed fairness goals that prioritized exposing content to followers first. PC1 felt that *"people who have already indicated interest in a creator's content—followers, liked— [should] be shown that creator's content as priority"* (PC1).

- ***Transparency.*** PC4 thought that fairness could be improved through greater transparency around *"policies, especially content revoking reasoning,"* while PC3 described that fairness could be improved through transparency about *"what factors into a successful post"* and *"algorithmic changes."*

Alias	Fairness Goal(s)	Fairness Definition	Fairness Metric
PC1	• *"Do not censor female/non male/non white/non cis bodies"*	*"Stop suppressing creators that aren't white men, women who show skin, or non-cis creators."*	Exposure Equality Metric
PC4	• *"Exposure consistency"* • *"Transparency around policies, especially content revoking reasoning"*	*"Providing equal opportunity for all users."*	Exposure Equality Metric
PD1	• *"Profiles have equal reach"* • *"Profiles have accurate audience"* • *"Profiles of marginalized identities not excluded from/within normative profiles"*	*"We will ensure that all profiles have equal visibility within the spaces/ groups/ populations the user desires to be seen with."*	Categorical Rejection Rate Metric
PD5	• *"Show a wide variety of individuals to mimic a real life setting"* • *"Show my profile to a wide variety of individuals"*	*"Mimic the outside world in its utopian state, exposing the user to all types of individuals to either match the user's existing perspective or to broaden it."*	Hotness Metric
PD7	• *"Users should be shown profiles (and have their profile shown) to a variety of people"* • *"Follow explicit user preferences, but not implicit ones (e.g. making assumptions based off of who the user has swiped on in the past)"* • *"Allow users to adjust their preferences and profile and adjust algorithm behavior accordingly"*	*"Follow stated user preferences while also allowing for diversity of shown preferences, and constantly adjust so as not to pigeon-hole any user."*	Diversity Metric

Table 4.1: Examples of fairness goals and their associated definitions and metrics designed by participants.

Dating App Users

Although transparency and exposure equality were also fairness goals for dating app users, the nuances of these goals differed in this domain.

- ***Exposure Equality.*** PD1, PD3, PD4, PD5, and PD7 all mentioned that every user should have an equal opportunity to be shown to prospective dates, regardless of their identity or dating preferences. PD3 additionally included that there should be equality in the allocation of benefits: *"every user should gain benefit from selection processes, showing and being shown to preferences."* PD2 also included a fairness goal to not promote *"any sort of racism, sexism, ableism, homophobia, transphobia, [casteism], and religious discrimination"* (PD2). This kind of exposure equality was also described as a *"utopian"* fairness goal because it would seek to *"expos[e] the user to all types of individuals to either match the user's existing perspective or to broaden it"* (PD5).

  In contrast, several participants described exposure equality as aligning someone's profile exposure with their explicitly shared dating preferences, such as PD1 who said *"I think like actually following through on... the settings that you put in and then actually like honoring those is like a very basic first step [towards fairness]."* PD7 also took this a step further and described a fairness goal where the algorithm should *only* take into account explicitly stated preferences, rather than implicitly observed ones: *"follow explicit user preferences, but not implicit ones (e.g. making assumptions based on who the user has swiped on in the past)"* (PD7).

- ***Transparency.*** Several participants detailed how improving transparency could increase the *feeling* of fairness through added agency on the platform, such as PD3 who said, *"transparency would go a really, really long way to making an app feel more fair to me."* This participant further shared that, *"getting a little bit more of*

a glimpse into how things work would go a long way toward making it seem more fair. ***Even if it isn't.***" This implied that even the *appearance* of transparency might make recommendation algorithms feel more fair based on the users' experience, regardless of whether or not the algorithm is *theoretically* fair. PD2, PD4, and PD5 also designed fairness goals related to improving transparency and agency surrounding why their profile is shown to others, and why certain profiles are shown to them.

In summary, although both content creators and dating app users developed fairness goals and definitions that promoted increased transparency and exposure equality, the nuances about which types of exposure would feel fair differed between these two recommendation domains.

Fairness Metrics

In the final activity of the focus groups, I asked participants to develop their own fairness metrics, based on their fairness goals and definitions. Each participant brainstormed what the goal of their metric was, the kinds of data they might need to measure this in practice, and the fairness "threshold" their metric would need to meet to consider the system fair enough for deployment. Here I describe several of these metrics for both recommendation contexts.

Content Creators

- ***Content Quality and Exposure Equality Metrics.*** Both PC3 and PC5 developed a metric to measure the quality of someone's content. These participants thought content exposure should only rely on the content *quality*, not on identity

attributes related to the creator. *"The algorithm shouldn't discriminate among factors that are not related to the content. For example, race and gender identity should not be a factor for content about art"* (PC5). To operationalize this metric, PC3 thought it would be necessary to collect demographic data (e.g., race, culture, language, gender identity, and personal aesthetics) to evaluate exposure differences between these attributes. This participant explained that if they were to use this metric, they would know that the platform is unfair if *"we still see small creators who make quality content getting very few views, seeing little to no growth in followers and engagement"* (PC3). PC4 developed a similar fairness metric that sought to measure equality of content exposure, based on demographic groups of content creators. They described that *"if all accounts [from different demographic groups] have a similar exposure percentage average (within 5% of each other), fairness is achieved"* (PC4).

- ***Popularity Bias Metric.*** Another metric developed by both PC2 and PC1 evaluated fairness for content creators with a small following. This metric is aligned with measuring the concept of Popularity Bias, the phenomenon in recommender systems where popular items receive most of the algorithmic exposure, while less popular items remain systematically under-exposed (Abdollahpouri and Mansoury, 2020). PC2 thought that this metric could measure what percentage of someone's social media feed is from profiles with small follower counts, and could be used to optimize exposure for those kinds of accounts. PC1's metric sought to improve popularity bias by measuring if certain content receives more or less exposure when posted from small accounts versus large accounts.

Dating App Users

- ***Transparency Metric.*** PD3 and PD4 developed a version of a transparency fairness metric. PD3's transparency metric was related to a questionnaire that users could fill out on a dating app. Their idea was that this metric could measure how compatible profiles are to one another and that fairness would be achieved if this information was shared with users. PD4's transparency metric instead focused on how much of the algorithms' functionality was being explained and shown to users—if every profile sorting and matching mechanism was being shown to users, this would be considered fair.

- ***Categorical Rejection Rate Metric.*** PD1 developed a fairness metric that would *"identify how many categorically incompatible profiles are being shown to a user,"* with a specific focus on improving compatible match performance for marginalized (e.g., LGBTQ+) users. PD2 developed a similar metric that sought to measure how much someone's profile exposure aligned with their explicitly stated preferences, specifically with respect to gender and sexuality. This participant determined that fairness would be "achieved" for this metric, *"if a provider's profile is presented to more than at least 50-60% of their intended target user"* (PD2).

- ***Diversity and Hotness Metrics.*** Finally, both PD5 and PD7 developed metrics to try to improve dating discrimination being perpetuated through dating apps. PD5 developed a *"hotness metric"* that sought *"to expose individuals to all different types of people with varying appearance [and] to not prioritize profiles that have received more likes/swipes or who appear more stereotypically attractive"* (PD5). PD7 developed a similar metric that sought to *"measure if the profiles being shown to a user are diverse across multiple features visible in their profile... within the user's stated explicit preferences"* (PD7). For this fairness metric, PD7 described that they

thought the platform would be deemed fair enough, *"if every user's recommended profiles score >=50% diversity."*

Who Should Design Fairness Metrics?

I also asked participants who they would ideally like to be involved in this process of designing fairness metrics for the recommender systems that they interact with. In all four focus groups, participants noted that the practitioners involved in this work should come from diverse backgrounds, which included a diversity of culture, geographical background, gender identity, sexual preferences, age, race, relationship status, and disciplinary background. PC1, PC3, PC6, PD2, and PD5 all specifically requested diversity for the programmers who might implement fairness metrics into the system. PD4 requested that fairness metrics should be designed by people who have used the apps before. PD1 requested that critical researchers, or *"people whose research focuses on marginalized identities"* be included. Finally, many participants (PC2, PC6, PC1, PC4, PD2, and PD6) also expressed that they would like fairness metrics to be designed with the *users* of these algorithmic systems.

Challenges with Measuring Fairness

During the focus groups, participants described various tradeoffs that might occur if their metrics were operationalized, as well as challenges that they faced when attempting to design fairness metrics generally.

Content Creators

PC2 shared that their goal to prioritize exposing content to followers might help content creators, but could also unintentionally lead to filter bubbles. *"It does kind of create a bit of a bubble... if you're seeing [just] the things that your followers are seeing"* (PC2). PC1 thought that their fairness metric (to measure what percentage of a creator's followers are exposed to their content) might only work well for content creators with a small number of followers, but not work as well for creators with large follower counts.

Another major challenge arose during a lively discussion between PC5 and PC6 when both participants realized that their fairness needs could not be met simultaneously on Instagram. PC6 described how, as an artist, they felt it would be fairer to ban AI-generated art from social media: *"I see the work of people I know who has been stolen into [AI-generated art] and these people are saying they're getting less and less views and that makes me very angry because obviously that's theft. And Facebook and Instagram are not making a ban on [AI-generated art], and that's really bad because I think they should be protecting [artists] ... and we cannot opt out of [generated content], which is really unfair"* (PC6).

In contrast, PC5, who creates AI-generated art, noted that banning or algorithmically suppressing that content would feel unfair to them. In this example, both PC5 and PC6 had different experiences of fairness and unfairness. Banning or suppressing AI-generated art might improve fairness for PC6 at the cost of fairness for PC5. Alternatively, exposing and recommending AI-generated art might improve fairness for PC5 at the cost of fairness for PC6.

Dating App Users

The main tradeoff that emerged for dating app users was a tension between increasing diversity of dating app recommendations while also preserving safety and explicitly stated preferences. PD7 described how increasing diversity of profile recommendations might also increase hate speech: *"there have been trans people who use [dating] apps and get matched with people who are transphobic and then like they get hate crimed on this app."* PD5 similarly expressed concern about increasing the diversity of recommendations on dating apps. They shared that this would be *"making the assumption that all people are kind and respectful and accepting, and it's just not the truth"* (PD5).

This led to a discussion about the responsibility of dating apps in general, where participants began to question if it is a dating app's responsibility to stop or hinder dating discrimination through fairness operationalization, even though it exists offline. PD5 thought it was the responsibility of dating apps to attempt to portray the world in its most utopian state. *"Let's portray a world that has like so much fairness, so much love. No racism ... like explosive diversity. I think ... it's the company's responsibility"* (PD5). PD6 similarly thought that it should be the responsibility of dating apps to not perpetuate any kind of discrimination or harm, but that certain apps could cater to specific preferences that already exist in the real world: *"A lot of Muslim girls I know use Minder ... people can have their biases and have a whole app for that. You know they can design for that in a non-harmful way for people who want to date certain types of people"* (PD6).

PD4 took this a step further and noted that *"there's a part of me that's like this whole [dating] process is inherently unfair, and that's sort of part of it. [Dating apps are] trying to make it too fair, [and] maybe [it's] not actually what people want"* (PD4). Ultimately, the participants agreed that this tradeoff would be inevitable when operationalizing fairness

for dating apps, and did not know how to remedy this challenge. *"A lot of times one of these things that could benefit one group can harm another group and it's hard to balance those"* (PD7).

Another set of challenges related to the efficacy of measuring fairness in practice. One example mentioned by PD1 and PD2 was the difficulty of deciding what their fairness "threshold" should be (how to determine which cutoff for their metric should be considered fair or unfair). *"What are the kind of arbitrary numbers that we're choosing to denote success or denote failure?"* (PD1). This participant also expressed concern that using fairness metrics in general might lead practitioners to believe they are measuring and optimizing for fairness when their measurements might not be accurately capturing users' lived experiences of fairness.

> *My concern... is that developers are going to take [metrics] as like the final step and they're going to kind of stop caring as long as they can keep this one certain numerical metric satisfied... they're not going to care to put resources towards more subjective, qualitative experiences of unfairness* (PD1).

PD3 added to this concern and felt like metrics might lead platforms to further exclude and marginalize certain users, just for the sake of good PR. *"If your populations are [small] enough... if you can make it as inhospitable as possible to the demographic that you're trying to have ... fairness for ... Then you're going to hit that [fairness threshold] every time, because there's no one there"* (PD3). All of these challenges highlight an important reflection about whether fairness ought to be measured at all. PD6 felt that fairness might be empirically immeasurable, even through proxies like fairness metrics: *"I thought [designing fairness metrics] was very hard to do. Thinking about how you can make these things empirical or like proving them empirically... it seems like an impossible task, honestly"* (PD6). I further explore these challenges of measuring fairness in the

following section.

4.3.3 Discussion

Opportunities and Challenges with Fairness Metric Design

In this focus group study, my collaborators and I noticed several opportunities and challenges of designing fairness metrics with providers from different recommendation domains and contexts. Below, I describe some of these observations.

Goals Versus Outcomes. As PD1 and PD3 described during their focus group, the *goal* to measure and optimize a recommender system for fairness might not necessarily guarantee the *outcome* of fairness for providers. This concern has been previously introduced through Goodhart's law—the notion that when a metric becomes a target, it ceases to be a good measure (Goodhart, 1975). In line with this challenge, PD6 also mentioned their concern that fairness measurement might become a PR goal rather than a real effort to improve users' lived experiences of fairness. This also introduces a challenge around fairness measurement generally—should recommender systems be attempting to operationalize a certain theoretical kind of fairness? Or should they be optimizing for users' lived experiences and perceptions of fairness? Previous work has shown that recommender systems can still be perceived as unfair by users, even when the generated recommendations are theoretically fair (Elahi et al., 2021). Thus, in future work, ML practitioners will likely need to discern what *kind* of fairness their system is attempting to measure, and what the end goal of that measurement should be.

Inherent Tradeoffs. Another major challenge observed during focus groups was the inherent tradeoffs that might exist when operationalizing certain fairness metrics over others. One example of this arose among the dating app users when the participants

expressed two competing desires: the desire to increase the diversity and exposure of their dating profiles (to decrease dating discrimination); and the desire to limit profile exposure to align with explicitly declared preferences (to decrease discomfort and unsafety on the platform). Both of these fairness goals were in direct conflict with one another, which could make it impossible to operationalize both goals at once. Another example of this kind of tradeoff emerged during the discussion between PC5 and PC6, where the decision to ban or suppress AI-generated content might make the platform feel more fair for some providers while making the platform feel less fair for others.

Domain Specificity. Throughout these focus groups, I learned that some fairness goals, definitions, and metrics were shared between content creators and dating app users. Both groups of participants were interested in improving transparency and exposure equality on their respective platforms. However, the nuances of how these goals might be measured or enacted differed between domains. For example, exposure equality for content creators PC1, PC4, and PC5 required that the algorithm prioritize recommendations based on the *quality* of someone's content. Previous work has shown that musicians also feel that not all music content is equally deserving of exposure and that algorithmic exposure should be based in part on quality (Dinnissen and Bauer, 2023). I note that exposing items based on their quality is an open challenge in recommendation research; although measures such as *expected exposure* from information retrieval attempt to capture this notion (Diaz et al., 2020), there are still many challenges with attempting to accurately measure item quality in practice. In contrast to content creators, exposure equality for dating app users PD1, PD3, PD4, PD5, and PD7 required that the algorithm give equal (and accurate) exposure to everyone, regardless of the "quality" of their profile. This difference implies that although different recommendation domains might have similar fairness goals, the operationalizations of these goals might need to be domain-specific.

Generalizing Between Domains. For recommender systems, increased item exposure

is sometimes thought of as a fairness guarantee (Abdollahpouri and Mansoury, 2020). However, by exploring providers' experiences of unfairness, I have learned that increased item exposure could, at times, actually lead to increased *unfairness* instead. Although several content creators in our study wanted more exposure of their content overall– previous work has shown that certain marginalized groups of creators do not want their content exposed to the "wrong" audience, for safety reasons. For example, DeVito (2022) discovered that trans creators wanted the algorithm to stop exposing their content to transphobic users because this could lead to hate speech. Similarly, in our study, PD1, PD2, PD3, and PD5 all shared that they did not want their dating profile exposed to the "wrong" audience, because it could lead to feelings of discrimination, exclusion, discomfort, unsafety, or loss of opportunity. These fairness goals, if shared between domains, could potentially be operationalized in similar ways.

4.4 Conclusion

The findings from both of these studies have significant implications for broader impacts and future work in this space. First, the participants (both end-users and providers) expressed a clear *interest* in being included in fairness design processes. While there is likely a selection bias in the participant population (e.g., individuals who volunteered for these studies may have a greater interest in fairness than the average stakeholder of recommender systems), this presents a valuable opportunity to engage stakeholders more systematically about their preferences for fairness in recommender systems. Of course, challenges remain—for example, individuals may express preferences that differ from their actual behavior in practice, achieving a representative sample of all stakeholders is difficult, and individual fairness perceptions may align more closely with personal satisfaction or self-interest rather than broader system-level fairness objectives. In addition to this, previous work has shown that people's perceptions of fairness are subject to change depending on several factors, including someone's gender, age, and *how* the system outputs are explained to them (Van Berkel et al., 2021). It is also well known that perceptions of fairness lie within the eyes of the beholder, which makes it incredibly challenging to operationlize in practice (Narayanan et al., 2024). These considerations underscore the importance of interpreting these study results with caution and situating them within the broader context of fairness research.

Nevertheless, these challenges can be mitigated through thoughtful approaches, and even imperfect stakeholder inclusion may be more beneficial than complete exclusion (for more discussion about how imperfect progress is better than no progress, see Section 6.3). Results from the focus group study highlight that fairness metrics derived from direct experiences of unfairness are often more closely related to addressing those experiences on platforms. This leads to an essential question: should we prioritize measuring and

improving fairness based on individual experiences of unfairness, or should we focus on theoretical, system-wide notions of fairness that may not always be perceptible to users? This tension between perceptions of fairness versus system-wide, theoretical notions of fairness is revisited in Section 6.2.2.

Moreover, qualitative research and stakeholder input have already been incorporated into responsible AI practices by some organizations, providing valuable precedents. For instance, OpenAI and Meta have both supported citizen panels and deliberative assemblies to address fairness and related challenges on their platforms (AI, 2024; Ovadya, 2021, 2023). These real-world examples demonstrate the feasibility and utility of involving stakeholders in addressing complex ethical issues (I do, however, note that these attempts to incorporate democratic participatory processes into technology development have fallen short in practice, something that I unpack in Section 6.2.5).

Even when fairness evaluations are not co-designed with impacted stakeholders, findings from the preliminary study in this chapter suggest that users may value awareness of fairness goals and interventions on the platforms they interact with, regardless of their inclusion in the design of those things. As a practical first step, platforms could introduce transparency around fairness efforts, which not only fosters user trust but may also enhance overall satisfaction with the system—offering a clear business case for such efforts.

In conclusion, while including end-users and providers in the fairness operationalization process may not always be pragmatically feasible, doing so can enhance user trust, agency, and satisfaction with the platform. It can also ensure that fairness evaluations more closely align with lived experiences of unfairness. At the same time, it is essential to acknowledge and navigate competing notions of fairness among stakeholders, as selecting particular fairness proxies may inadvertently privilege some stakeholders over others. For example, it is very likely that two dating app users might have different preferences for what they deem to be "fair" on the platform they use (as we saw in Study #2). So, deciding to

operationalize fairness for one users' ideal over another is an explicit decision that can lead to increased perceptions of fairness for some and not others. This dynamic is further explored in Section 5.2. Next, I shift the focus from users to another group of stakeholders who are critical to fairness operationalization: the practitioners.

Chapter 5

The Practitioners' Perspective

Overarching Research Question: What do practitioners know and need when operationalizing fairness in real-world recommender systems?

In this chapter, I describe the results from two interview studies with practitioners at Kiva (Smith et al., 2023c) and Spotify (Smith et al., 2023b). I explore what challenges practitioners face when conducting ML fairness work; what guidance might be needed when scoping their fairness goals, eliciting fairness considerations from stakeholders, and dealing with fairness conflicts among stakeholders; and how practitioners' platform design goals and decisions might differentially prioritize certain stakeholders over others. This research also assesses what kind of guidance practitioners might need when selecting appropriate fairness proxies and metrics for their given fairness scenarios to understand critical differences between fairness goals and their associated metrics.

5.1 Motivation

Although the perspectives of impacted stakeholders are an essential component of fair ML work, it is equally important to understand how practitioners operationalize fairness into their systems, and what guidance they need to do so effectively.

Challenges in Industry

Research in the ML fairness discipline has recently begun to explore how fairness work might need to be adapted to fit the constraints of real-world industry settings. Several studies with real ML practitioners in recent years have shed light on the difficulty that practitioners face when incorporating fairness into their existing practices (Rakova et al., 2021; Holstein et al., 2019; Richardson et al., 2021; Madaio et al., 2020). The general consensus among this research is that practitioners are largely lacking the specific guidance that is necessary to scope their fairness work, lacking the knowledge that is necessary to conduct fairness work, and lacking the tooling necessary to easily incorporate and scale fairness objectives in their large-scale systems.

Tooling in the context of this work includes open-source libraries that provide code to implement fairness metrics, as well as protocols, questionnaires, and worksheets to help educate and support decision-making in an algorithmic responsibility setting (Madaio et al., 2020; Cramer et al., 2018a; Moss et al., 2021; Bellamy et al., 2019; Xu and Doshi, 2019; Bird et al., 2020; Richardson et al., 2021). However, despite all of the available tools, practitioners still report a disconnect between these tooling efforts and what they actually need in practice (Madaio et al., 2022; Law et al., 2020; Veale and Binns, 2017; Lee and Singh, 2021; Richardson et al., 2021).

One major critique of ML fairness tools is that they do not provide the necessary guidance

for practitioners to *begin* exploring their fairness goals (Lee and Singh, 2021). Open-source libraries that provide metrics for practitioners are of little use if they do not know how to define the impact they want to measure, or scope their measurement context. Additionally, most ML practitioners are not trained in disciplines like ethics or philosophy (Saltz et al., 2019), which creates another barrier to entry for this complex decision-making space. Scoping fairness goals and constraints might seem very daunting without institutional or academic knowledge of disparate impact or fairness metrics. In response to these challenges, Saleiro et al. (2018) created a decision tree to help practitioners select an ML fairness metric. However, this decision tree assumes that the practitioner has prior knowledge of policy and ethics jargon, with some branches in the tree asking questions like, *"are your interventions punitive or assistive."* Additionally, this decision tree was designed for the context of binary classification, not ranking or recommendations. In the context of recommendation systems, fairness metrics and considerations are vastly different from a binary classification setting, especially since outcomes are not necessarily binary nor measurably favorable. There is rarely a "ground-truth" to compare the final recommendation lists against beyond assuming user engagement as a positive prediction (Beutel et al., 2019b). To combat some of these challenges, I aimed to iteratively create an easier-to-understand, recommender-specific fairness scoping framework, as showcased in Section 5.3.1, during Study #4.

Institutional Logics

Scholars in organizational studies widely concur that broader belief systems shape the way that stakeholders operate and make decisions within organizations; these values-driven organizing principles and patterns of practices are referred to as *institutional logics* (Thornton et al., 2012). For example, classes of fairness derived from the research

literature[1] serve as institutional logics of fairness when they serve as organizing principles for how individuals enact fairness. Numerous studies have characterized the ways in which multiple institutional logics co-occur within organizations—often in competition and contestation with each other (e.g., (Besharov and Smith, 2014; Greenwood et al., 2010; Pache and Santos, 2013; Reay and Hinings, 2009)). In the nonprofit context, for example, organizations are commonly required to negotiate the competing institutional logics that are part and parcel of being a mission-driven organization (e.g., prioritizing services to clients) and logics that are derived from their often-public sector funders (e.g., fiscal efficiency and accountability) (Evers, 2005; Binder, 2007; Mullins, 2006). As both a nonprofit organization and a financial institution responsible for managing individuals' investments, Kiva sits squarely at the boundaries of multiple institutional traditions with distinct institutional logics. As such, it is expected that multiple logics within this organization will manifest in different understandings of how to enact key values such as fairness.

Connecting this body of research to the domain of computing, Voida et al. (2014) further found that technologies also embody institutional logics in the ways they instantiate particular values. Even when organizational stakeholders agree on the importance of a particular value (such as fairness), that value may not be operationalized in technology in a way that is harmonious with stakeholders' various assumptions about and orientations towards how to put those values into practice. Such a mismatch creates tensions in practice that can create significant challenges for organizations and the clients they serve. This research, then, suggests that fairness-aware recommendation systems will need to be able to harmonize across multiple, potentially-conflicting logics about how the value of fairness should be operationalized.

[1] e.g., duty-based fairness or consequence-based fairness–see Section 2.2.5 for more information about these classes of fairness.

The Multiplicity of Co-Occurring Logics

Research has shown that not everyone thinks of fairness in a similar way (Narayanan, 2018; Smith et al., 2020). No fairness logics, then, are necessarily either *right* or *wrong*. Indeed, in nearly any context, the co-occurrence of multiple fairness logics is likely the norm and not the exception.

And yet, there has been relatively little recognition of the multiplicity of fairness logics in the ML literature (Smith et al., 2023b). Most existing research considers the differential impact of a system across only a single protected demographic class (e.g., race or gender). Even in cases where the impact of a system on multiple or intersectional groups is considered (e.g., (Hébert-Johnson et al., 2018; Kearns et al., 2018; Zhu et al., 2018)), the same concept of fairness is applied across all users, a decision that is at times justified through the application of Aristotle's concept of *Justice as Consistency* (Schauer, 2018), which requires that similar individuals are treated similarly (Dwork et al., 2012).

In contrast to the abundance of ML research that designs fairness interventions for only one class of stakeholders or applies only one concept of fairness, a subset of ML research has explicitly noted the benefits of combining multiple fairness definitions (Beutel et al., 2019a). In particular, multistakeholder recommendation system research has acknowledged and embraced the multiplicity of stakeholders and fairness concerns that often arise among different groups of stakeholders (Burke, 2017b; Abdollahpouri et al., 2020), such as tensions between consumer-side fairness (Ekstrand et al., 2018) and provider-side fairness (Sonboli et al., 2020c).

5.2 Study #3: Fairness at Kiva

The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation

(Smith et al., 2023c)

5.2.1 Introduction

In this work I explored the various kinds of fairness logics that arise in the domain of microlending recommendation using the platform Kiva as a case study. The Kiva organization is described in further detail in Section 3.1.1. Through analysis of interviews with 23 employees at Kiva, my co-authors and I uncovered four major fairness frameworks that were present when considering what the fairness goals of the Kiva organization are, and how those goals might impact the design of their recommender systems. The methods of this study are explained in more detail in Section 3.3.

Specifically, the research questions for this study were the following:

- **RQ1:** What fairness logics are present in Kiva's sociotechnical ecosystem of loan recommendation? Which stakeholders are prioritized by these different logics?
- **RQ2:** How might these fairness logics complement or conflict with one another?
- **RQ3:** In what ways do design interventions prioritize different fairness logics and/or different stakeholders?

The analysis revealed that participants' diverse considerations of fairness align with four different fairness logics: consequence-based, contract-based, character-based, and duty-

based. In the following section, I introduce these logics and explore where tensions arise between and among them. I conclude with design implications and methodological recommendations to assist practitioners with translating complex relationships among fairness logics into multistakeholder recommendation system design.

5.2.2 Results

> ***"What does it mean to be fair?"*** **(P21).**

Kiva's mission to improve global financial inclusion is, in part, motivated by global, systemic biases that create financial inequity, most often disproportionately for women and girls. The organization seeks to address this inequity by providing capital to women and other underfinanced populations (e.g., historically 80% of Kiva loans have gone to women). Participants' diverse accounts of how they enacted fairness in their work aligned with the four classes of fairness outlined in section 2.2.5: consequence-based, contract-based, character-based, and duty-based. In what follows, I characterize each class of fairness as it is enacted by Kiva employees—each logic of fairness serving as an organizing principle that shapes individual action and decision making. I note that these participants speak from the perspective of their individual experiences and roles at Kiva; they do not speak on behalf of the organization, itself. In what follows, I characterize each of these fairness logics from the perspectives of the participants.

Consequence-Based Fairness Logic

Considerations of consequence-based fairness focus on achieving the best *outcome*, which can be accomplished in two ways: (1) by maximizing benefit (e.g., by funding a greater number of high impact loans); or (2) by minimizing cost (e.g., by funding a greater num-

ber of low-risk loans). In general, consequence-based fairness logics consider stakeholders such as borrowers and lending partners—when maximizing benefit is equivalent to recommending more high-impact loans over low-impact loans; lenders—when minimizing cost is equivalent to recommending more low-risk loans; and the organization—when both maximizing impact and minimizing risk allows for more flow of capital, which enables the sustainability and growth of the organization's mission.

Benefit-Based Fairness Logic.

The first sublogic of consequence-based fairness is benefit-based fairness, where the underlying goal is to maximize benefit.

Maximizing Benefit By Increasing Impact. Participants used the word "impact" to describe how a loan might positively change a borrower's life. In this analysis, I use the word impact and benefit synonymously; while "impact" is used more often by participants, "benefit" is the language typically used in fairness literature. P2 noted that Kiva has created an "impact scorecard," which uses a set of metrics (e.g., the type of loan, the poverty index of the country, and the lending partner's historical impact) to determine a loan's likely impact. P5 mentioned that one of Kiva's goals is to increase funding for the most impactful loans; however, relying solely on an impact score might also be unfair in its own way:

> *"You have to choose one loan that's the most impactful compared to all other loans and it's really hard to do that because everybody needs help to some degree"* (P5).

Another way to increase impact is to increase funds flowing through Kiva to their borrowers: *"we obviously want as much money to go to these borrowers as possible"* (P8). The implication here is that in order to maximize benefit, Kiva should recommend loans that

lenders are most likely to fund over loans that are less likely to get funded.

Maximizing Benefit Through Purchasing Power. Another method for maximizing benefit is to recommend loans from countries where the value of the dollar is worth more. P4 suggested that some lenders prefer funding loans from low-income countries, believing that the money will go farther and have a greater impact: *"there is a perception that the $25 that you lend can go further for someone in another country which isn't wrong, but it's different"* (P4). P5 elaborated that the mentality of *"why [does Kiva give loans to individuals] in the United States if you can reach 50 more people in Africa for the average cost of the loan"* (P5) can bias investment towards low-income countries to increase the number of people served. P16 reaffirmed that people in high-income countries such as the United States also have issues with equality of opportunity: *"[there are] plenty of people in the US who also don't have equal opportunity"* (P16) and worried that lenders' preferences to maximize purchasing power by lending outside of the US could limit the chances of underserved people in high-income countries.

Cost-Based Fairness Logic.

A cost-based fairness logic strives to minimize the negative impacts of the lending process (e.g., if a borrower does not receive a loan). Most of the loans that are posted on Kiva's online marketplace have already been funded through Kiva's lending partners. Participants noted that lending partners depend on being recapitalized through Kiva's marketplace; if they are not, it will impact their ability to make future loans. While this will not affect the borrowers already featured on the platform (as they are already funded), it could affect the partner's decisions about which borrowers to finance with future loans.

Minimizing Cost By Minimizing Risk. participants most commonly suggested minimizing cost by reducing risks for lenders—recommending less risky loans. P8 explained that if new lenders are repaid, they will know that they *"did good"* because *"[the borrowers]*

were able to repay you" (P8). P3 added that recommending low-risk loans will not only keep lenders coming back to the site but will also help mobilize more capital for lending, generally—if a lender is paid back, they are more likely to reinvest that money in additional loans and the same dollar becomes a force multiplier for good. However, sometimes global and local issues can affect a lending partner's ability to repay a loan. P8 worried that over-focusing on risk and repayment rates might potentially lead to lower funding for borrowers living in countries that have volatile commodity markets. This participant added that focusing only on repayment capabilities might also only help the lending partners who are charging high-interest rates (which can hurt vulnerable borrowers). As a result, and in contrast to a cost-based fairness logic, P8 felt that it is okay to *"sacrifice a bit of your risk aversion for the sake of being more impactful"* (P8).

Contract-Based Fairness Logic

When participants' fairness logics aligned with contract-based fairness, it was always focused on enabling equality of opportunity—and nearly always for borrowers. Contract-based fairness aligns with Kiva's mission to improve global financial inclusion.

> *"The mission is to make sure those who are most financially excluded are the ones getting funded [...] Now, something that would be in line with our mission would be making sure those applicants are the first ones seen on the site as far as US loans go"* (P19).

P2 described this concept as *"filling the gaps"* in funding: *"it's not just that Kiva has less investments, it's that there are less investments in that country so that's where I want to focus. So for me, it's all about the gaps, where are the gaps in funding"* (P2). Several participants noted specific interventions to address these gaps, including recommending

loans for women and borrowers from low-income countries (P1, P2, P3, P4, P9, P12, P13, P22). P11 also mentioned interventions that are aimed at increasing inclusivity among borrowers—for example, providing translation services to non-English speaking borrowers and technical support for elderly borrowers.

However, enacting contract-based fairness is not as straightforward as it may seem. Despite being historically underfunded, on Kiva, female borrowers tend to be much more likely to receive funding when compared to their male counterparts (P3, P9, and P12). Similarly, P4 mentioned that small business owners in the US are also systematically underfunded in the online marketplace, raising the question of whether contract-based fairness applies to equal opportunity globally, or only within the context of Kiva's marketplace.

Character-Based Fairness Logic

Considerations of character-based fairness most typically explore the decision-making process that lenders go through when deciding who is "worthy" of a loan. Participants nearly exclusively discussed this logic while reflecting on the user experience of lenders and the ways in which borrowers are represented in the online marketplace. Participants believed that different lenders were more or less comfortable applying a character-based fairness logic and serving as an "arbiter" of which borrowers might be most "worthy:"

> *"There are some lenders who don't like choosing [which loans to fund], because they don't believe in themselves as an arbiter of who deserves. And there are other lenders who very comfortably sit in this place of evaluating who deserves a loan and who doesn't.... We see both of those mindsets play out in some of these situations where loan worthiness is considered more visible or obvious in certain contexts"* (P23).

P16 suggested that lenders on Kiva who have more microlending experience tend to prefer to make decisions about who to fund: *"There are some lenders who really care about microfinance and they came to Kiva ten years ago when microfinance was at its peak and they have specific ideas of what type of loan they're looking for" (P16).* However, P16 also shared that other lenders likely felt *"paralyzed by the choice"* when deciding who to fund; that they were not comfortable *"choos[ing] between two faces who [both] need help, and that doesn't feel good"* (P16). Similarly, P1 shared that they know *"friends and folks who've started using Kiva... [and] they struggle with selecting [a borrower] because they feel like they don't have enough criteria to pick like who's worthy"* (P1).

Given participants' inferences about lenders' comfort levels with a character-based fairness logic, many reflected on how to best support all lenders in a single online marketplace. If lenders want to make a decision about what borrower to fund using a character-based fairness logic, then Kiva might offer more information to support that decision making—but what information? While the microfinance enthusiasts mentioned earlier tend to want to see lending partners' risk scores, many lenders rely on visual information, especially the photo of the borrower (Jenq et al., 2015; Anderson and Saxton, 2016). While P3 explained that the borrower's photo is an important UX component (though not used algorithmically in any way to promote loans), it is also a place where discrimination and bias are very likely to enter the decision-making process: *"The best content really is the best pictures on our website, ... [but] what if [the borrower] didn't feel like smiling that day, what if culturally they're just less likely to smile in a picture, these are all problems... that I can't solve"* (P3).

For lenders who felt *"uneasy"* (P11) making decisions about who to fund (i.e., using a character-based fairness logic), participants mentioned the design of auto-lending features that would allow lenders to assign Kiva as a proxy for the decision. Participants also mentioned that minimizing the amount of information presented about borrowers might

also help these lenders, for example:

> *"A lot of lenders, probably a majority of lenders, do not care how we're doing this thing. They want to come to the website, find somebody whose story appeals to them, do a nice thing, and like move on with their day. They don't want to know the risk rating... they maybe don't even care what country, they just don't want all this information"* (P3).

Duty-Based Fairness Logic

participants' considerations of duty-based fairness align with the rules, duties, responsibilities, and accountability that Kiva has towards its various stakeholders. I recognize that duty-based fairness is more of a "meta" logic in that any rules that are chosen could overlap or conflict with any other fairness logic. However, this high-level process of deciding what an organization's values and duties are, and which rules they should follow to align with these values is an important part of enacting fairness at Kiva. I explore this complexity throughout this section.

P25 shared that *"one of our key core Kiva values is honor and integrity,"* which, for them, meant that Kiva has a duty to *"do the most right thing in the most right way"* (P25). But "doing the right thing" means different things for Kiva's many different stakeholders:

> *"So our algorithm is like determining whether or not people get air time, which, in a high supply environment where there are more loans than dollars, which is what we want... we want to maintain just a little more [loans] than we can fund.... We're responsible for who gets funded and who doesn't. We're creating the environment in which that happens, and there have been at different times like sort of like free market-y ideas at Kiva like "it's a marketplace, some*

> *stuff gets funded, some doesn't, it's not our fault," which is — it's completely not true. Like we're curating it to a huge extent, like the number of people involved, the number of rules, systems, processes involved in getting a loan onto the website is super high and the algorithm that presents them to lenders we made so while, yes, it's a marketplace and stuff, we are responsible for proper rulemaking"* (P3).

Different participants also prioritized different fairness logics when doing the "right thing." For P1, for example, doing the "right thing" meant maximizing contract-based fairness:

> *"If there's a way in which our machine is further perpetuating systemic racism or lack of access for certain people based on where they live or how they look or whatever, like those are things that we want to be aware of"* (P1).

For P2, doing the "right thing" meant protecting lenders' money, which meant maximizing cost-based fairness:

> *"Their focus is protecting lender money and protecting our repayment rate... trying to ensure that our lenders will get their money back, and that our repayment rate stays high and we're not losing too much money"* (P2).

These two operationalizations of duty—while both important and potentially complementary—define a tension or tradeoff in system design, a theme that I turn to next.

Tensions Between Fairness Logics

When multiple fairness logics co-exist, participants often noted tradeoffs or tensions between the logics and between the decisions that would result. P19 characterized these

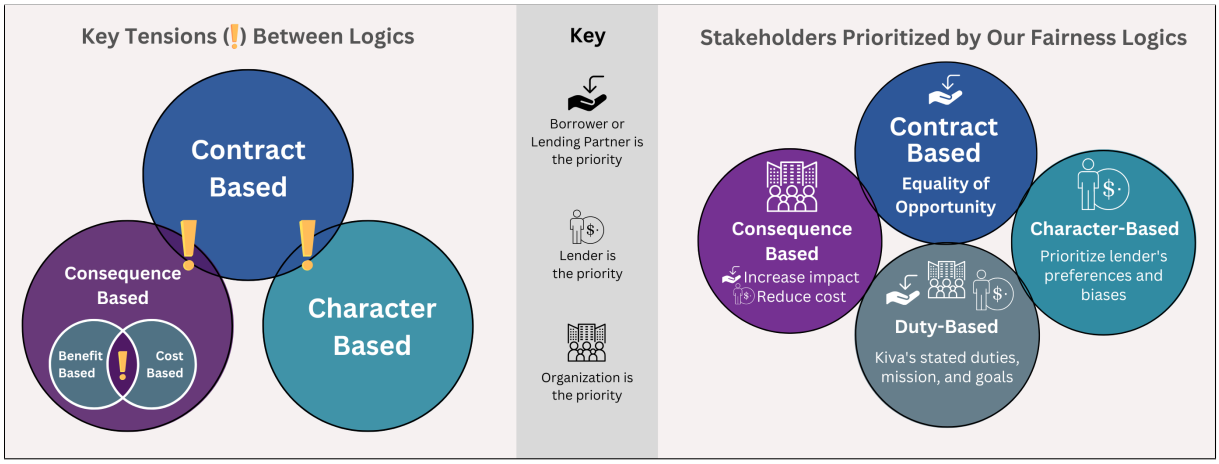


Figure 5.1: Left: key tensions between fairness logics. Right: the four logics and stakeholders that are prioritized by each. I note that duty-based fairness is excluded from the left side of this figure because participants were not in consensus about how to prioritize Kiva's duties in enacting fairness. Depending on which duties participants' emphasized, a duty-based logic could either conflict or complement the others.

tensions as a *"tough philosophical question"*:

> *"Then comes the question of like, is [intervening on loans] being equitable, and giving people like more opportunity who we feel are deserving of it? Or is that us kind of like playing God and not allowing the market to determine whether or not people should get funded like I don't know the answer to that question. That's a tough philosophical question"* (P19).

More specifically, in this section I identify three key tensions between fairness logics, as visualized in Figure 5.1.

Tensions Between Benefit-Based and Cost-Based Fairness Logics

The first tension occurs within the consequence-based logic, when a choice may be necessary between maximizing benefit (maximizing impact) and minimizing cost (minimizing risk). Several participants explained how, working from a consequence-based fairness

logic, it might not be possible to achieve both of these goals at once.

participants reported that Kiva has conducted internal research to better understand which kinds of loans have the highest impact on individuals and their communities. One thing that they learned is that many high-impact loans might also be high-risk, because *"moving [borrowers] out of financially vulnerable to financially stable"* (P12) sometimes requires lending to borrowers who are at a higher risk of not being able to pay the lender back.

However, prioritizing high-risk loans could reduce the flow of money through the online marketplace; if these higher-risk loans are not repaid, there will be less money for lenders to re-loan to other borrowers. Thus, a tension is created: if Kiva prioritizes maximizing impact, borrowers will benefit but lenders might not get repaid; if Kiva prioritizes minimizing risk, lenders will benefit, but certain borrowers might be systematically underserved—which is against Kiva's mission. Designing for consequence-based fairness, then, requires choosing how to balance these two fairness logics.

One participant felt that both of these goals (maximizing benefit and minimizing cost) could be accomplished simultaneously, by leaning on a very different operationalization of impact:

> *"There is this perception, I think a lot of the time, that risk and impact are kind of opposite sides of the spectrum. You can either be impactful or you can be low risk. I think that's a fallacy. I think they're kind of two parallel spectrums, spectra. Where you know you can be impactful, and you can be low risk at the same time"* (P8).

P8 gave as an example that if a lender were to lend $25 to a low-risk loan, then they are more likely to get paid back, and can consequently lend that same $25 to more loans

in the future—maximizing their overall impact while also minimizing their risk (P8). In this scenario, benefit-based and cost-based fairness logics complement rather than conflict with one another.

Tensions Between Contract-Based and Character-Based Fairness Logics

The second tension occurs between contract-based fairness and character-based fairness, when prioritizing equality of opportunity might conflict with the funding preferences of lenders. Participants described the challenges of improving equality of opportunity (contract-based fairness) while also attending to lenders who have their own preferences about who to fund or preconceptions about what kind of borrower should be funded (character-based fairness). As P1 summarized:

> *"[Fairness interventions] wouldn't be necessary if we didn't have some people who were deemed more worthy than others, and... those concepts do come from lenders"* (P1).

And P4 worried about relying solely on lenders to determine who is worthy of funding: *"The average person doesn't understand the impact of one specific loan"* (P4).

Numerous participants provided examples of how lenders' preferences have manifested in lending biases across different classes of Kiva borrowers:

- ***Types of businesses:*** P16, for example, has *"heard [lenders] say like 'I'm not going to lend to a beauty salon because that doesn't feel important to me."'* (P16). But making value judgements about what types of businesses are "worthy" or not imposes Western biases on borrowers. P21 provides the example of lenders not feeling comfortable lending to borrowers whose businesses would be illegal in the United States: *"if it's not illegal in the country of origin and it's not illegal due to*

international considerations, then what are our rights, from our point of view, to prohibit that type of lending" (P21).

- ***Gender:*** Many lenders prefer to fund borrowers who identify as female, systematically underfunding men on Kiva. P12 points out that in certain countries, men may actually need to be funded *more* than women, since they may be the only household member who provides finances for their family because *"gender roles are very different in different countries around the world"* (P12).

- ***Geographic regions:*** P9 noted that lenders also show regional preferences, for example *"Eastern Europe or certain parts of Asia"* do not fund as quickly as lending partners in Africa.

Each of these lender preferences creates a tension with ensuring equality of economic opportunity for Kiva's borrowers. P22 additionally worried that relying on lenders' preferences could unintentionally amplify lending biases on the online marketplace and result in underserved (long-tail) groups of borrowers:

> *"I think it's not all bad, but it is a cycle that we want to try to not perpetuate too deeply"* (P22).

And yet, pushing back on lender preferences too much might also have an adverse effect on equality of opportunity. As P11 explained, *"I think we found limits and how much we can like push things on people"* (P11). If lenders feel pushed too far and leave the platform, then equality of opportunity cannot be achieved, since there would be no funds left to give to borrowers.

Tensions Between Contract-Based and Consequence-Based Fairness Logics

The third key tension occurs between contract-based fairness and consequence-based fairness, when prioritizing equality of opportunity might be at odds with keeping enough money flowing through its online marketplace to keep the nonprofit afloat, pay back its lenders, and pay out to its borrowers and lending partners. P3, P11, and P25 all noted that these pragmatic needs of the organization (consequence-based fairness) all have to co-exist alongside Kiva's mission of increasing financial inclusion (contract-based fairness). Since a contract-based fairness logic prioritizes the "underdog," emphasizing this logic might, for example, entail nudging lenders toward higher risk loans than they might otherwise be drawn towards. P15 believed that it is risky to try to change lenders' behaviors in this way because you might be *"turning people off"* (P15) from lending, and doing so would conflict with the organizational need to keep money flowing through the online marketplace. P3 also worried that if lenders are pushed to fund high-risk loans but are not repaid, they might feel *"betrayed or misled,"* regardless of whether they were lending to a higher-risk borrower for the sake of improving financial equity.

So while working from a consequence-based fairness logic might be best suited to sustaining the organization, it might also reproduce financial inequality—which works against Kiva's mission. On the other hand, working from a contract-based fairness logic might place Kiva in a position where there are fewer lenders and less funding overall, putting the organization at risk. Balancing these tensions is a critical challenge for the design of microlending recommendation, which I discuss in the next section.

5.2.3 Discussion

In this research, me and my collaborators applied logics of fairness as an analytic lens to understand multiple, co-existing fairness considerations. We characterized four different fairness logics and identified three key tensions that arise at intersections between them.

These intersections between logics also represent sites of key design tradeoffs and opportunities to design for multiple co-occurring logics. In this discussion, I explore some of these design opportunities and design tradeoffs. Of particular importance here is the task of understanding how tradeoffs in design impact different stakeholders of Kiva's recommendation ecosystem.

Different fairness logics prioritize different classes of stakeholders, as shown in Figure 5.1. Organizing decision making from a contract-based fairness logic centered around Kiva's mission of financial inclusion prioritizes borrowers. Organizing decision making from a character-based fairness logic privileges lenders' preferences for lending and their perceptions about what classes of borrowers and loans are "worthy" to be funded. Organizing decision making from a duty-based logic impacts both lenders and borrowers/lending partners, as Kiva has a responsibility to all of these stakeholders. And organizing decision making from a consequence-based fairness logic prioritizes minimizing risks to increase the flow of capital through the online marketplace, benefiting the organization and the lender; it can also benefit borrowers if the emphasis is on maximizing impact; or all of these stakeholders if the loans maximize impact and minimize risk simultaneously.

In what follows, I explore two design cases, where tradeoffs exist between fairness logics and where opportunities for innovation can cater to multiple fairness logics at once. Many of these design options already exist—in varied combinations—across Kiva's online marketplace; here I detail them discretely in order to make explicit their relationships with fairness logics.

The Presentation of Loans

One tradeoff relates to what kinds of information Kiva might provide to lenders in the presentation of loan opportunities, their borrowers, and lending partners. As previously

mentioned, Kiva already provides a wide range of information about each loan on the platform, including:

- Basic information about the loan (e.g., total amount, expiration date, and what the loan will be used for);
- Qualitative and visual information about the borrower (e.g., a photo, their personal story); this design enacts a character-based logic; and
- Quantitative information about the borrower and their lending partner, including repayment rates, risk and impact scores; this design enacts a consequence-based logic.

Providing information like impact and risk scores can be helpful for certain lenders—those motivated by a consequence-based fairness logic—but can also perpetuate biases against lower-impact loans or higher-risk borrowers/lending partners, which undermines the contract-based fairness logic. Providing information about a borrower's personal life, such as if they are a parent, might help some borrowers and hurt others, leaving this design option open to introducing new biases for those working from a character-based fairness logic. One could combine multiple forms of information in the presentation of a loan in order to help organize decision making from multiple fairness logics. One could also personalize the genres of information on the presentation of loan page for different lenders based on what fairness logics they are inferred to enact when lending (i.e., engaging predominantly with information about a borrower (enacting a character-based logic), engaging predominantly with data about the repayment rates of lenders (enacting a consequence-based logic)).

Yet another design option would be to back away from interventions at the level of the individual loan and provide more generalizable information about *thematic loan categories*.

Kiva already provides options for lenders to filter loans categorically (e.g., "agriculture" or "education"). At this categorical level, one could present information about the importance of lending to each of these categories. For example, P9 pointed out that borrowers from certain regions could be systematically underfunded on Kiva. If a certain loan category is found to be underfunded, one could design a thematic cover page for all loan opportunities within that theme, with information about why this class of borrowers is important. This is similar to the "Spotlighted by Kiva" carousel that is already available on the platform, which highlights loans that are likely to not receive full funding. Another similar, yet alternative design could involve allowing lenders to choose to place a loan in this category, leaving the decision of what specific loan should be funded to Kiva (for lenders who are not comfortable enacting a character-based fairness logic) or providing additional context for lenders who want to select their own loan opportunities. This design would enact a combination of character-based, consequence-based, and contract-based fairness logics.

The Loan Discovery Page

As described above, Kiva currently uses a loan discovery page to help lenders find loan opportunities via a set of themed horizontal carousels, typically with 10–12 loans each. One set of design options revolves around choosing themes for these carousels, for example:

- Loans from categories, geographic regions, or borrower demographics that are the same as those the lender has lent to previously; this theme enacts a character-based fairness logic;
- Loans from categories, geographic regions, or borrower demographics that are more likely to go un-funded, regardless of lenders' funding preferences; this theme enacts a contract-based fairness logic; and
- Loans with high impact scores; this theme enacts a consequence-based fairness logic.

While all of these carousels could easily be included on the loan discovery page, the order in which carousels appear can also be an object of design. Their ordering could prioritize different fairness logics or could be personalized based on inferences about what fairness logics the lender has enacted through their previous loans.

Another set of design options revolves around choosing the subset of loan opportunities that are promoted on each carousel and the order in which they are displayed—their ranking—for example:

- Ranking well-funded ("popular") loans higher; this option enacts a character-based fairness logic, though it tends to reinforce lender biases;
- Ranking under-funded loans higher; this option enacts a contract-based fairness logic and mitigates the 'popularity bias' effect that can fall out of enacting a character-based logic;
- Ranking high-risk loans (less likely to get repaid) higher; this option also enacts a contract-based fairness logic but, instead, rebalances against some of the effects that can fall out of enacting a consequence-based logic; and
- Ranking low-risk loans (more likely to get repaid) higher; this option enacts a consequence-based fairness logic.

However, as multiple participants in our study shared, if one goes too far in *"push[ing] things"* (P11) on lenders, the sustainability of the online marketplace is put at risk. As a result, one might also want to explore ways of designing the loan discovery page—both the carousels and the loan rankings within each carousel—to prioritize multiple of these co-occurring fairness logics at once and experiment with weighting them in different ways in different contexts (Liu et al., 2019; Sonboli et al., 2020c). Social choice theory, for example, has been successfully used as a framework for algorithm design that combines

multiple fairness logics (Sonboli et al., 2020b; Burke et al., 2020, 2022a). One might also personalize the design approach by inferring which fairness logics lenders have previously enacted in their lending decisions and weighting their recommendations based on a similar balance of fairness logics (e.g., (Liu et al., 2019; Sonboli et al., 2020c; Li et al., 2021b; Farastu et al., 2022)).

Selecting *which* logics to prioritize in *what contexts* on Kiva's online marketplace and *how* to operationalize these logics is a compelling challenge for future work.

5.3 Study #4: Fairness at Spotify

Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective

(Smith et al., 2023b)

5.3.1 Introduction

In this work I explored the challenges that practitioners face when scoping their fairness goals and determining how to empirically measure those goals. Through a semi-structured interview study with 15 ML practitioners at Spotify, my co-authors and I iteratively co-designed a decision-tree framework that helps practitioners select which fairness metrics to use based on their fairness goals in a recommendation setting. The methods are described in further detail in Section 3.4. The full decision-tree framework can be found in the Appendix in Section 7.4.

Specifically, the research question for this study was:

What guidance do ML recommendation practitioners need when scoping fairness objectives and identifying related fairness metrics?

The main contributions of this work are threefold: (1) a literature review of recommendation and ranking fairness metrics, which I coalesced and categorized into Table 5.1 (consumer-side fairness metrics) and Table **??** (provider-side fairness metrics); (2) a preliminary decision-tree framework organizing said literature to help practitioners scope fairness objectives and identify fairness metrics; and (3) results from semi-structured interviews with practitioners.

Similar to previous observations on fairness in recommendation systems in practice (Beattie et al., 2022), me and my collaborators found that there are no standard decision-making processes for practitioner teams to approach operationalizing fairness metrics in recommendation systems. We also learned that practitioners need more guidance when defining fairness objectives and scoping those objectives into their associated metrics. I discuss how to address these findings based on the current state of research on fairness in recommendation systems and suggestions from participants in our study.

5.3.2 Results

In this section, I describe the challenges explicitly mentioned or implicitly observed while practitioners attempted to scope fairness-related conceptualizations for their selected recommender system. I compare these challenges to those identified in previous research and explore the implications of our results.

Scoping Qualitative Harms

Research literature popularly defines impact in terms of "fairness", but I found that this wording may not adequately reflect the various impacts practitioners need to evaluate in an industry setting. This finding was uncovered while discussing the results of the scoping worksheet with each participant to understand which potential impact they would like to use while navigating the decision tree. I found that some had encountered challenges when deciphering the ambiguity of the term "fairness" while scoping their fairness problem.

Interview participants expressed having difficulty knowing how to define "fairly" in a quantifiable way. P14 described "fairly" as *"legalese,"* , noting that we would need a concrete definition of fairness, such as *"in proportion to the opportunity available to [providers],"* in

order to capture this concept accurately. P13 similarly described that they would need a definition for "fair," ideally based on market research and in line with company values and business goals. This could include a conceptualization of, for example, concrete algorithmic impacts rather than 'algorithmic fairness'. This suggests that it could be helpful for practitioners to relate the choice of impact conceptualization, as well as metrics, to specific company and stakeholder goals or values, which has also been discussed in previous work (Madaio et al., 2022; Cramer et al., 2018b). This also points to the need for dedicated business support, which is often more feasible in larger corporate settings, leaving smaller organizations at a disadvantage due to less access to resources (Beattie et al., 2022).

Navigating the Decision Tree

Here I share the challenges participants faced when navigating between decision-making nodes and branches in the decision tree. The full decision-tree framework can be found in the Appendix in Section 7.4.

The Multistakeholder Constraint Branch

Beyond scoping qualitative harms and their associated fairness objectives, as noted above, participants struggled to understand common nuances used in fairness research literature to define various fairness constraints. When evaluating the multistakeholder constraint branch, some participants were unsure if they should explore fairness metrics for providers or consumers. Even though the decision tree offered examples of each stakeholder (e.g., providers could be content creators, consumers could be end-users), some of the practitioners remained unsure of which stakeholder they should prioritize for measuring fairness.

Another example of this difficulty in choosing between stakeholders arose during P5's interview, who indicated a potential challenge determining if a system's recommendations

reinforced country stereotypes. During this interview, P5 assumed that a consumer fairness metric would be most appropriate, since stereotypes might immediately impact those who consume the recommended content. However, they also noted that provider-fairness metrics might be equally appropriate for their context, especially if providers were being stereotyped due to their country of origin. P4 experienced the same challenge. They noted that fairness concerns could impact both providers and consumers simultaneously because they are *"two sides of the content ecosystem"* (P4). This challenge aligns with previous work (Madaio et al., 2022) that highlighted difficulties practitioners experience when deciding which stakeholder population to prioritize for fairness interventions in a multi-stakeholder environment. This suggests that practitioners may need more guidance when selecting which stakeholders to prioritize when conducting fairness interventions, particularly in scenarios where design tradeoffs may be necessary (Burke, 2017a).

The Group vs. Individual Constraint Branch

Participants also had difficulty navigating the branch addressing the individual and group fairness constraint. Interestingly, a lack of knowledge about the difference between individual and group fairness was sometimes accompanied by a heightened level of confidence about this choice. For example, P2, P5, P8, and P14 all reported that one of the easiest decisions for them to make during the interview was whether they were measuring group or individual fairness—even though *all* of these participants chose an option for their specific fairness context that appeared less aligned with literature definitions. I categorized these participants' "more aligned" options as metrics that matched the population of users they had chosen for their context. For example, P5 specified that they wanted to measure the impact of stereotypes on providers between different countries. In this case, it would be more "aligned" to define the provider country as a group type and measure group fairness. However, P5 chose to measure individual fairness instead.

For fairness problems that aligned with group provider fairness metrics, the decision tree

prompted participants to choose whether they were interested in measuring fairness for one group at a time or multiple groups at a time. This decision also appeared challenging for some participants. For many fairness problems, both kinds of group fairness metrics could be applicable. For example, P8 wanted to measure whether certain long-tail creators were being unfairly represented in recommendation lists. This participant described that they could measure fairness by comparing their long-tail artist group's exposure against a target distribution (one group at a time) or by comparing their long-tail artist group's exposure against another long-tail artist group (multiple groups at a time). P8 said that both of these approaches might be equally appropriate, which made it difficult for them to know which branch to choose. This experience again reaffirmed our suspicion that there may be multiple, equally appropriate metric categories for a given fairness problem; which might add more difficulty for practitioners when deciding which of these many metric(s) to implement.

System Component Branch

The challenge of discerning nuances in fairness measurement terminology persisted when practitioners were asked to choose between system components for measurement. This branch splits based on the need for ranking or distribution metrics (e.g., measuring fairness of item exposure across all recommendation lists versus measuring fairness of item exposure in the top-k ranking of a specific recommendation list). At this point, the framework asks practitioners if they are more concerned with "poor representation in any group's item distribution or candidate generation list," (distribution metrics) or if they are more concerned with "any group being low in rankings or engagement for top-k or candidate generation lists," (ranking metrics).

P4, P10, P13, and P14 all stated that this choice between distribution and ranking fairness was the hardest decision they had to make in the decision tree. Both P4 and P13 were unsure which option to choose, since their fairness problem could apply to both categories.

P4, P10, and P14 shared that this decision was difficult as they felt they did not have the specialized knowledge they needed concerning metrics that would fit their recommendation system. For example, these three participants disclosed that they did not understand some of the more technical vocabulary used to describe distribution versus ranking metrics. This suggests that education efforts might be best focused to target the gaps in practitioners knowledge: for some education about technical jargon and system components; for others education about fairness terminology.

The Fairness Objective Constraint

When participants did not have the necessary knowledge to make an informed decision on which fairness objective to finally choose, they sometimes chose the "easiest" path. For example, one participant navigated through the decision tree by selecting the fairness options and paths that were most familiar to them, even if they might have been missing out on metrics that were better aligned with their given scenario. This participant further described that the "easiest path" for them was one that tried to optimize "utility" in the system, because that path aligned with terminology also used for business goals, and it *"would be easier to get PM buy-in if needed."* This suggests that practitioners may choose metrics that are easy to understand, implement, or explain using existing business terms or perspectives, which may be the main priority to move the needle. Previous work has also described this phenomenon, where standard or well-known fairness definitions are adopted and assumed to be applicable across contexts (Selbst et al., 2019).

Note that such considerations are valid—a very complex metric that is hard to understand but technically correct is unlikely to change decision-making, however, it is essential to clarify the consequences of such choices to assess whether a chosen metric is an appropriate proxy. This challenge could also be attributed to our choice of binary options at several of our decision-making nodes. Future tool iterations may benefit from allowing users greater flexibility in choosing from multiple possibilities.

Challenges Beyond the Framework

Throughout the interviews, I observed that there is not always a "correct" path or metric. Rather, there might be many equally-appropriate ways to measure fairness for a given scenario, all of which target evaluating different metric categories. When encountering these challenges, I asked participants to describe *who* they would ask for decision-making guidance. I found that participants were uncomfortable making fairness-related decisions on their own, especially when they faced challenges deciding on constraints.

Some of the participants expressed a desire to involve a diverse group of roles in these types of conversations, such as domain experts, model owners, product managers, engineers, user researchers, and business leaders. This was shown by P10 and P13, who both wanted to involve a *"fairness expert"* to help them feel confident in navigating this area. Beyond a fairness expert, P13 wanted to involve a user researcher to help them refine the population for fairness analysis as well as business leaders to help them align fairness tasks with business goals. P14, a user researcher, wanted guidance from more quantitative data-oriented counterparts such as data engineers, data scientists, and analytics engineers in order to understand *"the limitations of the model [because] they know what the data that we collect is actually capable of capturing"* (P14). P13 also further described that even though it would be nice to make these decisions with a group of fairness experts, they also did not want to exclude anyone who had little knowledge of fairness but was still interested in participating in the decision-making process.

Decision-Making and Implementation Gap

During our interviews, some practitioners who focused more on implementation indicated a need to involve other disciplines for fairness decisions in theory, but were less optimistic about the extent to which colleagues could be helpful in practice. For example, P9 described that these scoping decisions should ideally involve a combination of individual

engineers or product developers working with a PM. However, in practice, P9 found that even though the PM should be involved, in a prior workplace they had experienced that fairness decisions ended up coming down to individual engineers implementing metrics. P5 had a similar experience to this. They described that from their perspective, ideally, fairness rationale and a more specific problem statement would come from a PM, but more detail would be necessary to implement than what would likely be provided. This points to the gap between conceptualization versus more detailed choices necessitated during implementation and a need for more clarity in roles and expectations in this process.

As seen with many other participants, P7 (who is an ML-focused engineer) was struggling to choose between the distributional and ranking metric constraints, and stated that part of their difficulty making this choice was because *"this is not just an ML problem, but also a product decision"* (P7). They said that in practice, they would probably go to their PM to help them refine their fairness goals, and then would bring the fairness scenario up with their entire engineering team. Then, they would probably try prototyping some of the metrics, and would check their results with a data scientist. They described the whole process as, *"a lot of different decisions made by different people, together and separately"* (P7). This description of fairness work as a collaborative, interdisciplinary process has also been explored in previous work (Passi and Barocas, 2019; Passi and Jackson, 2018), where the authors describe how decisions in practice are often made by a combination of different actors, such as data scientists, PMs, and user researchers, and are decided through collaborations and negotiations. Similar to that previous work, both P7 and P10 described their ideal collaborative process as one that highlights everyone's area of expertise.

These participants described that PMs would be the most helpful for high-level fairness decisions (e.g., the stakeholder constraint or the fairness objective constraint), while people like ML engineers could be responsible for the more low-level implementation decisions

(e.g., the system component constraint). Previous work has also highlighted that practitioners are uncertain about *who* should be doing ethics work, and where the responsibility for doing ethics work should fall within an organization (Metcalf et al., 2019). Organizations might benefit from providing central guidance, including clarity on roles and responsibilities, while future academic work or case studies could explore the consequences of organizational choices.

5.3.3 Discussion

Challenges Beyond the Framework

Throughout the interviews in this study, I observed that there is not always a "correct" path or metric. Rather, there might be many equally-appropriate ways to measure fairness for a given scenario, all of which target evaluating different metric categories. When encountering these challenges, I asked participants to describe *who* they would ask for decision-making guidance. I found that participants were uncomfortable making fairness-related decisions on their own, especially when they faced challenges deciding on constraints.

Some of the participants expressed a desire to involve a diverse group of roles in these types of conversations, such as domain experts, model owners, product managers, engineers, user researchers, and business leaders. This was shown by P10 and P13, who both wanted to involve a *"fairness expert"* to help them feel confident in navigating this area. Beyond a fairness expert, P13 wanted to involve a user researcher to help them refine the population for fairness analysis as well as business leaders to help them align fairness tasks with business goals. P14, a user researcher, wanted guidance from more quantitative data-oriented counterparts such as data engineers, data scientists, and analytics engineers in

order to understand *"the limitations of the model [because] they know what the data that we collect is actually capable of capturing"* (P14). P13 also further described that even though it would be nice to make these decisions with a group of fairness experts, they also did not want to exclude anyone who had little knowledge of fairness but was still interested in participating in the decision-making process.

Decision-Making and Implementation Gap

During our interviews, some practitioners who focused more on implementation indicated a need to involve other disciplines for fairness decisions in theory, but were less optimistic about the extent to which colleagues could be helpful in practice. For example, P9 described that these scoping decisions should ideally involve a combination of individual engineers or product developers working with a PM. However, in practice, P9 found that even though the PM should be involved, in a prior workplace they had experienced that fairness decisions ended up coming down to individual engineers implementing metrics. P5 had a similar experience to this. They described that from their perspective, ideally, fairness rationale and a more specific problem statement would come from a PM, but more detail would be necessary to implement than what would likely be provided. This points to the gap between conceptualization versus more detailed choices necessitated during implementation and needed clarity in roles and expectations in this process.

As seen with many other participants, P7 (who is an ML-focused engineer) was struggling to choose between the distributional and ranking metric constraints, and stated that part of their difficulty making this choice was because *"this is not just an ML problem, but also a product decision"* (P7). They said that in practice, they would probably go to their PM to help them refine their fairness goals, and then would bring the fairness scenario up with their entire engineering team. Then, they would probably try prototyping some of the metrics, and would check their results with a data scientist. They described the whole

process as, *"a lot of different decisions made by different people, together and separately"* (P7). This description of fairness work as a collaborative, interdisciplinary process has also been explored in (Passi and Barocas, 2019; Passi and Jackson, 2018), where the authors describe how decisions in practice are often made by a combination of different actors, such as data scientists, PMs, and user researchers, and are decided through collaborations and negotiations. Similar to that previous work, both P7 and P10 described their ideal collaborative process as one that highlights everyone's area of expertise.

These participants described that PMs would be the most helpful for high-level fairness decisions (e.g., the stakeholder constraint or the fairness objective constraint), while people like ML engineers could be responsible for the more low-level implementation decisions (e.g., the system component constraint). Previous work has also highlighted that practitioners are uncertain about *who* should be doing ethics work, and where the responsibility for doing ethics work should fall within an organization (Metcalf et al., 2019). Organizations might benefit from providing central guidance, including clarity on roles and responsibilities, while future academic work or case studies could explore the consequences of organizational choices.

Table 5.1: Consumer-side fairness metrics with their associated categories and constraints as depicted in the decision tree. Some fairness metrics were assigned names if they were not explicitly named in the associated paper.

Fairness Category	**Fairness Constraints & Objectives**	**Metric(s) and References**
Individual Calibration	Recommendations should match consumers' previous interests.	Calibrated Fairness (Steck, 2018)
Utility vs. Merit	System should distribute utility for consumers (individual or groups) based on their merit or need.	Generalized Cross Entropy (Deldjoo et al., 2022)
Group Utility	System should distribute utility equally between groups of consumers.	Non-Demographic Fairness (Hashimoto et al., 2018), Average Accuracy (Ekstrand et al., 2018), Differential Satisfaction (Mehrotra et al., 2017), Pairwise Accuracy (Beutel et al., 2019), epsilon Fairness (Li et al., 2021a)
Group Error	System should have similar error rates between groups of consumers.	Value Unfairness (Yao and Huang, 2017), Absolute Unfairness (Yao and Huang, 2017), Underestimation Unfairness (Yao and Huang, 2017), Non-Demographic Fairness (Hashimoto et al., 2018)
Group Identity	Recommended items for consumers should not be related to group identity.	Non-Parity Unfairness (Yao and Huang, 2017), Discriminatory Skew (Ali et al., 2019), Pairwise Accuracy (Beutel et al., 2019), Counterfactual Fairness (Li et al., 2021b)

Table 5.2: Provider-side fairness metrics (part 1) with their associated categories and constraints as depicted in the decision tree. Some fairness metrics were assigned names if they were not explicitly named in the associated paper.

Fairness Category	**Fairness Constraints & Objectives**	**Metric(s) and References**
One-Group Representation	Provider group representation in recommendations should match a pre-defined baseline distribution	Minimax KL Divergence (Das and Lease, 2019), Average Provider Coverage Rate (Liu et al., 2019)
One-Group Rank Utility	Groups of items should be ranked according to their utility for consumers.	Equal Expected Exposure (Diaz et al., 2020), Inter-group Pairwise Accuracy (Beutel et al., 2019)
One-Group Rank Proportions	Top-k ranking should contain a specific proportion of items from a specific provider group.	Skew at K (Geyik et al., 2019), Top-K Fairness (Asudeh et al., 2019), Normalized Discounted Difference (Yang and Stoyanovich, 2017), Viable-Lambda Test (Sapiezynski et al., 2019)
Multi-group Item Exposure	Provider groups should have an equal item exposure distribution.	Minimax KL Divergence (Das and Lease, 2019)
Multi-group Item Relevance	Provider groups' probability distributions, preference ratings, exposure, etc. should be proportional to their items' relevance.	Kolmogorov-Smirnov Statistic (Zhu et al., 2018), Absolute Difference in Mean Ratings (Zhu et al., 2018)

Table 5.3: Provider-side fairness metrics (part 2) with their associated categories and constraints as depicted in the decision tree. Some fairness metrics were assigned names if they were not explicitly named in the associated paper.

Fairness Category	**Fairness Constraints & Objectives**	**Metric(s) and References**
Multi-group Rank Proportion	Top-k ranking should contain a specific proportion of items from provider groups.	Normalized Discounted KL-Divergence (Yang and Stoyanovich, 2017), Normalized Discounted Ratio (Yang and Stoyanovich, 2017), Viable-Lambda Test (Sapiezynski et al., 2019), Ranked Group Fairness (Zehlike et al., 2017), Rank Parity (Kuhlman et al., 2019)
Multi-group Rank vs. Relevance	Provider groups' item rankings should be equally proportional to their items' relevance.	Disparate Impact Ratio (Singh and Joachims, 2018), Demographic Parity Constraint (Singh and Joachims, 2018), Disparate Treatment Ratio (Singh and Joachims, 2018), Listwise Fairness (Zehlike et al., 2017), Inter-group Pairwise Fairness (Beutel et al., 2019)
Under-Over Exposure	Provider groups and/or items should not be systematically under-exposed or over-exposed in recommendations.	Gini Index (Vargas and Castells, 2014)
Individual Rank vs. Relevance	Item clicks, exposure, ranking, etc. should be proportional to item relevance for similar items.	Equal Expected Exposure (Diaz et al., 2020), Disparate Treatment Ratio (Singh and Joachims, 2018), Disparate Impact Ratio (Singh and Joachims, 2018), Demographic Parity Constraint (Singh and Joachims, 2018), Listwise Fairness (Zehlike and Castillo, 2020), Equitable Individual Amortized Attention (Biega et al., 2018)

5.4 Conclusion

The findings from this chapter highlight several key implications for fairness evaluation practices and future work in industry settings. For Study #4, we learned that there is no single "correct," "ideal," or "optimal" fairness evaluation. Once fairness goals are selected, there are many ways to measure them in practice, each involving different stakeholders, systems, proxies, and metrics—all of which could be equally valid depending on the context. This underscores the importance of providing guidance to practitioners when conducting fairness evaluations.

Such guidance cannot prescribe exact actions (given the multitude of equally valid approaches) but can help practitioners weigh the trade-offs, benefits, and consequences of their decisions. Providing structured reasoning frameworks for these evaluations could be more beneficial than attempting to deliver definitive answers.

This principle also applies to the Kiva study (Study #3). Numerous ways to incorporate fairness into Kiva's platform design emerged, each potentially benefiting different stakeholders. Consequently, the scoping decisions faced by Kiva practitioners were not about *whether* to incorporate fairness but rather about *how* to do so—acknowledging that some stakeholders might benefit more than others from specific operationalizations of fairness.

This means that fairness evaluations must navigate competing interests among organizational teams and external stakeholders regarding what constitutes fairness. It is unlikely that any fairness operationalization will satisfy *all* stakeholders of a system, including users, practitioners, and the organization as a whole. This is why I believe that creating structured categories or lists of fairness considerations and offering prioritization guidance for practitioners will be critical.

Central to this work is the need to justify fairness-related decisions made during the scop-

ing and design process. Transparency about these decisions—and their consequences—is needed. I also believe it is essential for practitioners to explicitly acknowledge that fairness interventions will benefit some stakeholders while disadvantaging others. Therefore, I recommend practitioners avoid implicitly optimizing specific fairness objectives without openly addressing the inherent trade-offs involved in their operationalization.

The findings from Study #4 also emphasize the interdisciplinary and cross-functional nature of effective fairness evaluations. Fairness experts are needed to define goals and identify appropriate proxies based on the data and context. Modeling teams must assess data availability and system constraints, while evaluators develop and measure fairness proxies accurately. Domain experts contribute knowledge about stakeholder considerations, PMs make final decisions and accept responsibility for outcomes, and qualitative researchers ensure assumptions align with the lived experiences of impacted stakeholders. This collaboration mirrors calls for interdisciplinary, cross-functional, and cross-team collaboration in prior fairness research (Ali et al., 2023; Madaio et al., 2024).

Additionally, fairness can be integrated into recommendation ecosystems in various ways, including through dataset creation, constraints on recommendation lists, re-ranking methods, and UI or platform design (e.g., carousels highlighting underserved providers). The ways we evaluate and enhance system fairness can align with any of these strategies. Practitioners need clearer guidance on these options and support in determining which are most suitable for their context and goals.

The overarching takeaway from these studies is that fairness is not a one-size-fits-all concept. Practitioners most need guidance on scoping the journey from: "I want my system to be fairer," to: "this is the kind of fairness I care about, this is how I will measure it, and these are the stakeholders we will prioritize with this evaluation." This critical early-stage decision-making process, moving from fairness aspirations to operationalization, is where guidance is currently most lacking and where future work should prioritize its efforts.

Specifically, there is a need for tools and frameworks that focus less on prescriptive advice and more on descriptive guidance–helping practitioners navigate the key scoping decisions, understand the trade-offs inherent in their choices, and evaluate the broader impacts of their fairness operationalizations on diverse stakeholders. By developing such resources, we can empower practitioners to create systems that are not only more equitable but also deeply aligned with the contexts and communities they serve.

Chapter 6

The Experts' Perspective

Overarching Research Question: What are potential best practices for operationalizing fairness in real-world recommender systems, keeping in mind the practical constraints of this setting?

6.1 Motivation

There are competing ideologies in the ML fairness research community as to whether fairness work ought to focus more on ideal, theoretical conditions of fairness, or more practical (albeit imperfect) conditions (Kommiya Mothilal et al., 2024; Wilkinson et al., 2023). While research has proposed numerous fairness definitions, metrics, and methods (e.g., (Narayanan, 2018; Hutchinson and Mitchell, 2019; Chouldechova and Roth, 2020; Mehrabi et al., 2021)), the practical realities and consequences of applying these concepts in industry settings create a challenging evaluation landscape for practitioners. Industry practitioners often face competing pressures, including aligning fairness efforts with business objectives, navigating resource constraints, and addressing limited access

to stakeholder perspectives (Rakova et al., 2021; Scheuerman, 2024; Ali et al., 2023; Holstein et al., 2019; Madaio et al., 2022). These challenges demand a nuanced approach to fairness that acknowledges and works within real-world limitations.

Fairness evaluations in industry typically follow three stages of decision-making: (1) scoping the evaluation, (2) measuring fairness, and (3) interpreting results (Barocas et al., 2021; Smith et al., 2023a; Beattie et al., 2022), as shown in Figure 6.1. However, these stages are complex and require extensive expertise in ML fairness, as well as interdisciplinary and cross-team collaboration (Ali et al., 2023; Madaio et al., 2024).

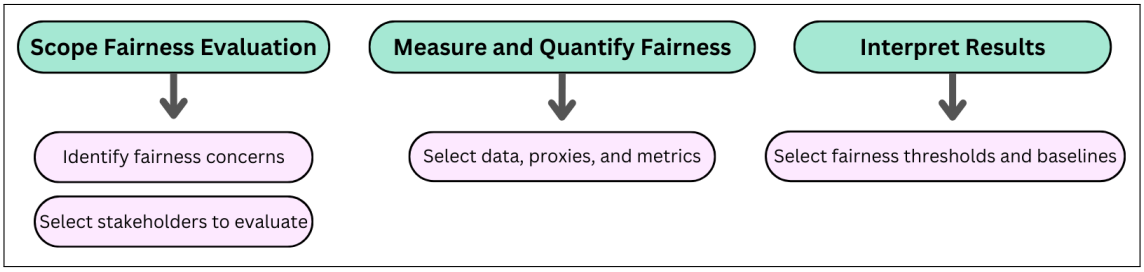


Figure 6.1: The different decision-making stages for designing and conducting ML fairness evaluations in real-world settings.

The Constraints of Industry

The main goals of machine learning in industry are to build, ship, update, and maintain products and systems (Scheuerman, 2024; Ali et al., 2023). It has been shown that research insights from fairness or responsible AI evaluations need to support product or company goals in some way (Scheuerman, 2024), and work in this setting is always motivated by and rewarded based on metrics tied to business objectives (Ali et al., 2023).

However, it can be challenging to prove fairness' impact quantitatively, which can hinder practitioners' ability to link responsible AI efforts to the organization's success (Rakova et al., 2021). Previous work has also shown that resources for fairness evaluation are only made available when business imperatives or goals align with this work (Rakova et al.,

2021; Holstein et al., 2019), and so incentives for ML fairness tend to work best when they come top-down from leadership (Holstein et al., 2019). ML fairness in industry is primarily motivated by PR concerns (Widder et al., 2023; Madaio et al., 2022) and regulations (Rakova et al., 2021), even though previous work has shown that many individual practitioners would like to incorporate fairness into their ML systems if given the time and incentives to do so (Madaio et al., 2020).

In addition, fairness work in industry often follows the approach of *"move fast and break things"* (Holstein et al., 2019), and is thus *"optimized for speed"* (Madaio et al., 2024). Real-world systems and product teams are also incredibly complex and frequently shifting, which means that fairness work in practice also needs to consider the dynamic structure of these systems and their associated development teams when designing and conducting fairness evaluations (Beattie et al., 2022; Holstein et al., 2019; Ali et al., 2023; Madaio et al., 2024).

Finally, there are noted power imbalances in industry settings, which means it is not always easy to speak out about ethics or about how fairness work should be conducted (Widder et al., 2023; Ryan et al., 2024; Widder et al., 2024; Madaio et al., 2020). Thus, practitioners might not have the opportunity to make decisions about how fairness evaluations are designed or conducted, and are at times required to perform evaluations within very specific constraints and bounds over which they have no control.

Beyond the practical constraints of the industry setting, there are many other noted challenges that practitioners face when conducting ML fairness work, as described at length in Section 2.3.

Theories of Change in ML Fairness Evaluation

How then, can we evaluate ML fairness given these challenges, business constraints, and theoretical hurdles that render the problem impossible to solve holistically? The economist Amartya Sen contrasts two ideas of justice, captured by the Sanskrit terms *niti* and *nyaya* (Sen, 2009). A focus on *nyaya* tries to envision the ideal conditions for perfect justice in a theoretical sense. Sen critiques Western moral philosophy for its emphasis on this idealized pursuit. He contrasts it with the concept of *niti*, comparative improvement of justice, which he argues can be sought and achieved without reference to a perfect ideal and to which a perfect ideal may be little guide. I take inspiration from this argument by exploring how ML fairness evaluation can be improved without demanding that these practices match up with a theoretical ideal. This perspective aligns with the notion of *reform*, or working within the system to improve it, rather than *abolition*, seeking to dismantle the system to improve it; both perspectives are of interest to the ML fairness research community (Wilkinson et al., 2023).

The Futures Cone, introduced by Dunne and Raby (2024), and shown in Figure 6.2, provides a framework for conceptualizing futures across a spectrum of probable, plausible, and possible scenarios. While the model encourages imagining alternative and ideal futures to critique and expand our designs, industry settings often constrain fairness work within the boundaries of what is probable and plausible due to business limitations. In this chapter, I adopt this framework as a foundation to argue for a pragmatic approach to fairness evaluation in ML systems–one that acknowledges these constraints while emphasizing actionable strategies within the probable and plausible domains. Although this approach prioritizes feasibility over idealism, it allows practitioners to make incremental progress toward fairness, recognizing that engaging in imperfect, pragmatic work is preferable to inaction. Pushing the boundaries of what is plausible and possible for ML fairness evaluation in industry is also crucial; I discuss potential avenues for this in Section 6.2.5.

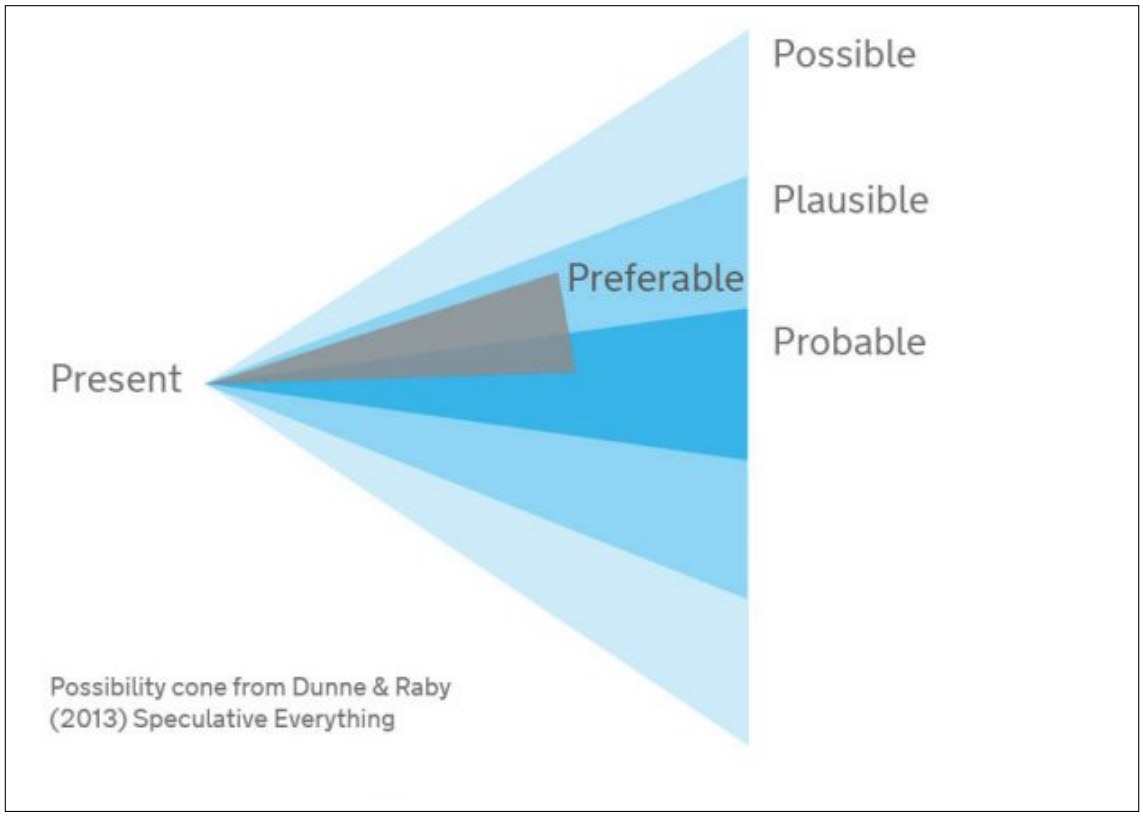


Figure 6.2: The Futures Cone from (Dunne and Raby, 2024).

6.2 Study #5: The ML Fairness Experts' Perspective

Pragmatic Fairness: Evaluating ML Fairness Within the Constraints of Industry

6.2.1 Introduction

Through a a semi-structured interview study with 21 ML fairness experts, I explored the complex problem space of ML fairness evaluation in industry, using recommender systems as a domain to investigate how practitioners address these constraints and what opportunities exist to advance fairness under practical conditions. The methods for this study are described in further detail in Section 3.5.

During this study I explored the following research questions, all within the context of evaluating real-world recommender systems:

- **RQ1:** How do ML fairness experts conceptualize fairness, and what goals do they associate with its evaluation?
- **RQ2:** What constraints arise in ML fairness evaluations within industry, and how do they influence decision-making?
- **RQ3:** How do experts think fairness evaluations can be improved within the practical constraints of industry?

The main contributions of this work include: (1) an exploration of how experts conceptualize ML fairness and its evaluation goals; (2) introducing **pragmatic fairness** as a concept; (3) providing examples of how pragmatic fairness is conducted; (4) providing examples of how we can improve pragmatic fairness; and (5) an exploration of the limitations of prag-

matic fairness and how the research community can enhance efforts to evaluate fairness in a satisfactory, though imperfect, manner.

6.2.2 What is Fairness and What is the Goal of Evaluation?

I begin by grounding the results of this work by investigating how participants in this study conceptualized "fairness" within the context of recommender systems, as well as the overarching objectives of conducting ML fairness evaluations in real-world applications.

In the literature, the overarching goal of incorporating fairness into machine learning systems is to justifiably allocate the benefits and harms of a system among its stakeholders (Crawford, 2017; Shelby et al., 2023), usually while avoiding disproportionate impacts on groups and individuals who are already likely to experience societal marginalization.

In general, since recommender systems are multistakeholder, so are fairness conceptualizations in this domain. Related to the previous literature on multistakeholder fairness in recommendation systems (Ekstrand et al., 2022), several participants in this study also differentiated between **provider-side** fairness and **consumer-side** fairness. P7 defined provider-side fairness through two considerations: (1) how often a provider's items show up in recommendation lists; and (2) how high these items are ranked in those lists: *"I think the problem is... mostly about visibility.... at the end it's all about whether someone or something makes it to the final [recommendation] list, and if they do, where in the list they appear."* P8 developed this definition further by explaining how we can perform disaggregated evaluations across groups of providers' and their item distributions to measure provider-side fairness: *"is the content ... fairly distributed across a certain parameter? [Are] our evaluation metrics fairly distributed across [providers]?"* (P8).

In contrast, P7 explained how consumer-side fairness is about making sure there is no

unwanted demographic bias in the recommendations that a consumer sees. *"You could also have fairness from the [consumer's] perspective, like what type of content I am seeing versus what type of content you are seeing. If what I'm seeing I think has some bias in terms of... let's say, gender or age... certain recommendations that might not be fair to me... from the user perspective, it's also whether what I'm seeing as a user is reflective of my taste, regardless of certain... sensitive features"* (P7). Similarly, P8 shared that consumer-side fairness could also be focused on ensuring there is equality within the user experience of the recommendation ecosystem: *"you want to make sure that when you're doing these types of evaluations where you're testing different systems against one another, [you're checking] whether [the system works for] every single age group or for every single gender, or for different levels of expertise... To an extent, that is also fairness, right? It's a much higher-level fairness. It's the fairness of the user experience... that is the user centric fairness question"* (P8).

Beyond formalizing fairness through the lens of impacted stakeholders, I also learned from the participants that there are two high-level conceptualizations of fairness that can be operationalized in systems: (1) theoretical, system-level fairness; and (2) lived experiences of fairness. Here I describe both of these fairness conceptualizations and their limitations.

Theoretical, System-Level Fairness

Many participants defined fairness through the lenses of equality, equity, diversity, and bias. These conceptualizations of fairness involve thinking of fairness as high-level, system-wide constructs that we can quantitatively measure holistically for many users at a time.

P5 described fairness as *"everybody [getting] what they deserve,"* but then proceeded to describe that this was *"very complicated and ultimately politically loaded because... people have different ideas of what people deserve and people have different ideas of who is*

deserving and why. So that's the challenge." This participant described that one way to mathematically confront this challenge could be to find an optimum between equality of opportunity and equality of outcome.

> *"So one of the principles that I really try to get across to people is you need to separate predictions and policy... What are the properties of your prediction? What kind of ... errors does it make? Where do you fall on the [pareto] frontier between equality of opportunity and equality of outcome or other trade-offs? There's this whole set of questions that are, in a sense, purely technical or I mean – to use a dirty word – purely objective in the sense that you're trying to match an observable reality"* (P5).

This distinction between opportunity versus outcome (also known as **equality versus equity**) was also named by P14, who shared that *"there are political traditions that hold that fairness is everybody getting the same, and there are political traditions that hold that fairness is giving people what they need... and so like many systems before them, the values of the developers kind of get baked in and the values of system developers in academia in the United States skew a little bit toward... the 'fairness as justice' perspective... equity... where people get what they need. And so you see that lens a lot in places like FAccT, but it's not necessarily true in industry"* (P14).

Another theoretical formalization of fairness shared during interviews was the concept of **diversity**. For example, P11 shared that for them, *"fairness is also similar to diversity. [I] want to [expose] as many [items] as possible... In my research, I want to [expose] diverse opinions on the same topic and surface them to the top results for different people... We want minority sellers, we want the new emerging sellers or brands to be [exposed] to the users, we want to give some chance to those tail items"* (P11).

P12 also described how fairness in recommendations could be related to diversity, and

shared how this formalization can actually be good for online marketplaces: *"if you define fairness with respect to diversity, then we know that diversity is definitely very useful in many, many cases... For any kind of marketplace you do need that diversity... Because obviously just because you like one genre that doesn't mean that's what you're gonna watch or listen to all the time... So, if you think [of diversity] as a mapping to fairness, then yes... people are already always interested in fairness" (P12).*

This participant also described how fairness could be related to **bias**, and how this formalization allowed for easier statistical evaluations. *"Most times we have associated fairness with the notion of bias, and basically saying that a fair system is the one that lacks bias.. to be honest... the nice thing about that is it allows us to more systematically do experiments. So the nice thing about... bias [is that] you can measure [it] quantitatively... you can use statistical parity or you can use exposure or group size, all kinds of things... And if your fairness is strictly associated with [bias], you can say... the system is fairer than this system or we did this thing and it has a fairer ranking or recommendation"* (P12).

P11 echoed this, and shared that they also think of bias as a more objective measure of fairness: *"personally I think that bias is more like objective. You can actually measure... like quantitatively say this bias is towards this population"* (P11).

However, P12 also critiqued this approach as *"naive,"* in part because it was unclear to them if bias represents how *"how fairness plays out in the real world."* (P12). I return to this concept of fairness in the real world shortly in Section 6.2.2.

P12's critique of bias was not the only critique that participants shared about mathematical formalizations of fairness. P9 wondered if *"our notion of fairness mathematically correspond[s] to the way normal people think about fairness,"* and claimed that *"if we optimize for some academic notion of fairness that does not correspond to the way the community itself thinks about fairness, then what have we really done?"* (P9). P20 similarly shared

their *"anthropological critique of fairness"* that calculating fairness mathematically is *"not a realistic approach to the experience of fairness and the social reality of using a recommender system and being affected by one"* (P20). This pariticipant elaborated further on this idea by describing that mathematical definitions of fairness are *"all approximations of a variety of values that people have that are not held as mathematical definitions... People do not hold 'rigorously provable' as ideals in their heart"* (P20).

In addition, P13 felt that attempting to quantify fairness was almost like trying to *"chase this number, whatever the number might be, without necessarily asking if that is the main problem [you] might face on their system"* (P13). This participant felt that trying to achieve a certain numerical threshold of fairness could distract practitioners from focusing on other, more salient types of harm in their system. P4 also shared their opinion that *"mathematical formulations [of fairness] are great in gambling games and game theory applications"* but they *"just haven't found them to be that useful when you try to implement [fairness metrics] in practice"* (P4). This critique was bolstered by an example of recommendations on dating apps, where there is a notion of popularity, merit, and relevance that fairness metrics struggle to take into account. This participant felt that *"you could come back and come up with any mathematical model of fairness, but if you don't inject the reality that there are some people who are immensely popular, and by virtue of that popularity, have leverage in getting to choose... who they match with... then [fairness metrics are] not going to work"* (P4).

Alongside these critiques, P8 also pointed out that mathematical fairness is necessarily reductionistic, and often requires us to choose arbitrary values for metric heuristics that have *"no link to any real world value."* P12 similarly felt that measuring fairness forces us to *"sacrifice that realism in a way,"* and shared that even if we are optimizing for some mathematical notion of fairness, there is still the limitation that we are likely not actually improving *"individual or societal or group fairness in the real world"* (P12). However, even

though P12 felt that there is a *"disconnect between what's realistic and what's scientific,"* they did share that several benefits of mathematical formalizations of fairness, including that it allows us to *"clearly define metrics,"* to *"measure things," "report things,"* and to more easily *"compare things"* (P12).

Even though mathematical formalizations of fairness allow us to more easily evaluate system-level fairness, P8 described their concern that if this fairness operationalization was not perceptible for users, then there might be no point in doing it.

> *"One thing that I think we're not doing enough in recommender systems research in general is thinking about things in terms of discernibly, right? We're optimizing algorithms and... to the extent where we're [improving] performance by like 0.01%, what kind of impact does that have on real people? So I'm almost thinking like, if you have that difference in NDCG, my question almost becomes...**does it matter?** Like can you actually measure an observable difference in user experience? Whether that is click through behavior or... perceived recommendation quality, or even system satisfaction, can you actually link those differences to differences in those values?" (P8)*

In line with P8's critiques, I now turn towards these user values, satisfaction, and perceived recommendation quality, and explore what this kind of fairness could look like in real-world recommender systems.

Lived Experiences of Fairness

In contrast to theoretical, or mathematical notions of fairness, many participants described how fairness could be operationalized through the lens of individual's lived experiences and perceptions of what they deem to be fair–to try to more accurately capture an in-

dividual's perceptible experience of fairness. For example, P19 shared that *"[fairness is] an abstract concept... if you ask what is your lived experience with a system–like what is your lived experience using autocomplete on your phone... what is your lived experience with obtaining loans... rather than asking about fairness... It's about asking about lived experience versus asking opinions about an abstract concept"* (P19).

Several participants described that they thought perceptions of fairness was a more meaningful construct to evaluate, in part because it is actually perceptible to the users. For example, P9 shared that one problem they had struggled with was whether their *"notion of fairness mathematically correspond[ed] to the way normal people think about fairness"* (P9). P21 similarly liked *"the idea of figuring out what is actually meaningful to people and their lives, and then thinking about... things to measure to help us improve [our fairness metrics] based on that"* (P21). One example of how to evaluate fairness based on users' perceptions was provided by P9, who described the idea of a **just notable difference**.

> *"This idea of a 'just notable difference'... people may not notice the difference between like a .05 and a .08 in some unfairness metric... but they would notice... 0.05 versus 0.25. That's very noticeable in the data... They might notice 'hey, compared to my friends, I'm getting much [better] results'. So thinking about that as a standard rather than what is this threshold for good or like what is the relevant amount of change [or] minimal increase [we want] in a model?"* (P9).

P8 described how you can measure individuals perspectives of fairness by asking users for their *"gut feeling"* about whether or not they think the system is fair for them.

However, one participant critiqued this framing of fairness, and shared how they thought that *"users feeling like they're being treated fairly is not the same thing as fairness... It's user satisfaction* (P20).

Continuing this critique of lived perceptions of fairness, P3 described how you might create a theoretically equitable system that is still perceived as unfair by individual users, only because it does not match their preferences. P3 shared an example in the context of Google search results:

> *"For example... [it] might be that I don't want to hear any news from right-leaning news websites, and every time I [use] Google, if those things show up, I get frustrated and I don't want to use Google. But the ideal fair weight would be to... weight left news sites and right news sites equally, so that no matter who uses Google, it's always the same mix at the top. This would be fair from a holistic population standpoint, but it would be unfair if you scope your definition of customer experience to [an] individual basis. Because all of a sudden, if I really don't want any right-wing news websites, I can no longer change or specify that... It has made that decision for me. And then that's where I feel it actually becomes unfair"* (P3).

P7, P8, P11, and P13 described that people's perceptions of fairness are incredibly subjective and might differ by age, culture, gender, and other demographic axes. P7 additionally felt that people's perceptions of fairness will often be biased to benefit themselves, rather than historically disadvantaged groups of people, as is the case with theoretical fairness. *"Often people don't have a fair judgement of themself. So if I'm a niche provider and I have got 1000 impressions, I might not be happy with this even though I know... that's exactly how much I deserve. But I just want to express my negative opinion just to say you need to change your algorithm as a tool to give feedback... [because] maybe if they change the algorithm I will get more [impressions]"* (P7).

P18 described how this tension between perceptions of fairness and system-level, theoretical notions of fairness might be because system-level fairness requires us to have

aggregated metrics about groups of people to determine how fair the system is, something that we cannot identify from individual experiences.

> *"This is a challenge... individual perception and the system level effects may not be aligned very well... Many concepts of fairness are not very perceptible from the individual perspective. Procedural types of fairness often are, like I understand the process by which the system decided to recommend my stuff, and I think that is treating me fair in a process kind of way. But process fairness doesn't necessarily provide any guarantees about, say, the overall equity of the overall distribution of resources... That's a phenomenon that really only arises over, say, a group of providers in aggregate, or a group of people in aggregate. So there's a fundamental challenge here. What we're trying to optimize... is often not something that very cleanly maps to anyone's experience. So it's hard... there can be a bunch of different goals."* (P18).

P13 shared that another challenge with incorporating lived experiences of fairness is that it can be hard to determine if someone's experience of unfairness is a *"one-off"* versus a replicable and repeated *"statistical bias"* within the system as a whole.

All of these results reinforced that fairness definitions, goals, proxies, measurements, and evaluations are not straightforward in real-world settings. The concept of "fairness operationalization" could mean a variety of approaches and outcomes depending on the interpretation of fairness from the practitioners themselves. The choice of whether to evaluate the system for an imperceptible, though theoretically justifiable, notion of fairness versus users' perceptions of fairness is one that practitioners must make. This decision highlights that there is no singular "right" or "wrong" way to evaluate fairness; rather, each choice involves tradeoffs and carries distinct consequences for the system's various stakeholders.

The Goals of ML Fairness Evaluation

Barocas et al. (2021) previously described three considerations when determining what the goals of an ML fairness evaluation should be. First, is the intention of the evaluation to demonstrate performance disparities between groups or to uncover potential causes of performance disparities? Second, is the evaluation focusing on previous disparities that have happened to specific people, or focused on potential future disparities that could affect a generalized group of people? And third, is the evaluation intended to be confirmatory (to provide conclusive evidence) or exploratory (not intended to provide conclusive evidence)? Beyond these categorizations, the participants in this study shared several, more nuanced, goals that could emerge for fairness evaluations in practice.

First, one goal of fairness evaluation shared by participants was to start conversations about what could be wrong with a model/system to assess whether interventions might be necessary. P4 said that *"if you don't measure it, you don't know if it's a problem"*– implying that fairness evaluation is one step towards identifying if certain problems or harms actually exist in the system. Similarly, P13 shared that *"if you don't know if your system is flawed, you cannot even know how to fix it."* P9 felt that fairness evaluations were a useful way to *"start a conversation about whether there are some performance issues or fairness related performance issues in the model."*

Another goal of fairness evaluations shared by participants was to work towards creating a more equitable system overall for different stakeholders. For example, P18 said in the context of provider-side fairness, *"one significant goal is like, as we're building and deploying these systems, are we contributing to a world in which a broad set of people from a broad set of backgrounds with various different interests and visions etcetera, are able to have equitable access to customers and listeners and audience and all of those things that the recommender mediates access to?"* (P18). In contrast, this same participant described

that for consumer-side fairness, *"under serving a group of users is a market vulnerability"* and so *"from a consumer side, we can look at our [fairness] goal [as] we don't want to be undeserving some of our users"* (P18). Similarly, P12 said one of their goals for fairness evaluation was to *"make sure that nobody is harmed in an obvious way*, so even if the system still had certain harms, those were not *"because [of] certain demographics... it's just a characteristic of [the] overall system... [So] practically you kind of start defining it in that context, as opposed to societal fairness... No group gets special treatment. No groups get [differential] harm... Let's aim for that"* (P12).

Another goal described by participants was a duty towards trying their best to make sure their system did not amplify certain inequalities in the world over time, such as P10 who said one of their goals with fairness evaluation was to not make *"the existing bias worse"*, or P20, who shared that they thought there might be an *"an obligation, like an ethical obligation to recommend"* less popular providers in recommendations, and that fairness evaluations are one method to help us achieve that goal. P16 also described how even though users might not always be able to perceive certain fairness operationalizations, *"it's important to still not feed into the system... [because] things get worse over time. Like sure, it might not matter now, but in 50 years. Like could we have done something to make a minute difference?... [Like] the butterfly effect"* (P16).

These varied conceptions of ML fairness and its evaluation goals underscore the complexity of this space for both academics and industry practitioners. The term "fairness" itself is highly fluid, taking on diverse meanings when applied to the evaluation of machine learning systems. With this foundation, I now turn to examining the practical constraints that industry practitioners face and how these further shape and influence ML fairness evaluations.

6.2.3 The Path From Constraints to Satisficing

In this section I describe three constraints that heavily impacted the participants' decision-making when designing ML fairness evaluations in industry settings: (1) balancing competing interests; (2) lacking power and/or access to necessary supports; and (3) prioritizing decisions based on what could enable buy-in from leadership and/or the organization. I describe how the participants navigated these constraints by **satisficing**—seeking satisfactory rather than ideal outcomes (Simon, 1978), and explore the nuances of this approach.

Constraint #1: Balancing Competing Interests

One unique aspect of doing fairness work in industry is that it requires balancing competing interests between the fairness evaluation team and the product/modeling teams, which involves fitting in with pre-existing workflows, timelines, and employee bandwidths. Several participants shared that fairness work could not happen within industry without getting support from product/modeling teams. For example, P13 said, *"if you do not own the model... there's not much you can do without having the support of the [product] team."*

In industry, fairness evaluators are often separate from modeling teams. P9 described how modeling teams have *"very different incentives"* than fairness auditors, and are highly motivated to ship products quickly, which may conflict with fairness auditing that can slow production down. P9 noted that this *"very big conflict of interest"* is one benefit to keeping the fairness evaluation team separate from the modeling team. P13 also shared their experience that sometimes modeling teams might say they are interested in fairness evaluation, but that they *"do not have the bandwidth this quarter... [so] please come back in three to six months,"* which P13 said, *"is totally reasonable."* P2 also noted that

fairness work often depends on *"being opportunistic,"* because resources like data contracts or budgets may disappear unexpectedly.

Fairness work also faces competing interests between stakeholders, such as consumers and item providers. P12 explained that increasing recommendation diversity (one potential heuristic for fairness) could sacrifice personalization accuracy, which is crucial for engagement and revenue. Similarly, P20 noted that prioritizing less popular providers might conflict with business imperatives to generate revenue, leading to decisions that *"might have no reason to be aligned with fairness."* These challenges are consistent with previous research, which found that business imperatives often drive fairness decisions (Madaio et al., 2022).

Moreover, improving fairness for one stakeholder group might reduce fairness for another. For example, P11 explained that in a music recommender system, balancing personalization for end-users with fairness for providers–such as giving new artists a chance to be recommended–requires careful consideration. As P11 noted, *"we have to strike a balance between different parties, fairness, and also the system performance."* P12 similarly pointed out how balancing fairness for end-users and advertisers is crucial, stating, *"what we are looking for is a good balance where we can retain [both] the users and the advertisers."*

Constraint #2: Lack of Power and/or Access to Necessary Supports

Another constraint of ML fairness in practice that is fairly well-known is that those who conduct fairness evaluations in practice are often not those who have decision making power at an organization (Metcalf et al., 2019). For example, P7 explained how as a fairness evaluator, they might not have any power to decide which fairness definitions are evaluated in the system, because *"each country [has] specific laws for what's considered*

underrepresented groups or what's considered harm, so those definitions may most likely sometimes come from... higher level... as [an algorithm designer or evaluator], you might not decide on the [fairness] definition, because that's given to you." Or, as P15 described, *"it is more often than not sort of a top down decision... maybe a project manager or product manager is defining what that [fairness] success criteria is."* This lack of power could limit practitioners' abilities to influence and potentially improve fairness evaluation approaches.

Some of the participants reported that beyond a lack of power, practitioners also often lack access to necessary supports for fairness evaluations. For example, it is common that practitioners lack the demographic data they might need to conduct disaggregated fairness evaluations (Andrus et al., 2021). P9 and P13 shared their challenges conducting evaluations in industry without the demographic data they would need to perform disaggregated evaluations for group fairness. P13 said that without the data they needed, they *"had to always come [up with] or find more creative ways to do this kind of analysis."*

In addition, P17 shared how they would love to base their fairness evaluations off the real experiences of impacted users, but in practice they do not have access to these people or this data. P15 said that even though it would be ideal to elicit fairness concerns from users, in practice this was not practical because asking users about their experiences was too big of a PR risk, especially if users were being asked to provide opinions about unreleased products or features, this could be too big of a risk for the company even if users were required to sign NDAs. P10 also shared how they did not have access to resources to elicit perspectives from enough users to get a meaningful sample to generalize their experiences into larger system-level evaluations: *"[small user studies] would just [have] so much variance in it that we would not be able to draw firm conclusions."* P3 similarly felt it would be ideal to get feedback about fairness from users, but realistically, they *"can't deploy [the model] on 4000 people from every country in the world within two weeks before*

it has to ship, and so for practical reasons, oftentimes [the volume of] these user studies... is incredibly small", so *"if you want to do a good user study, it's really hard."*

Constraint #3: Getting Buy-In

Another essential component to enabling fairness work as described by the participants was to get buy-in from leadership, so that this work is mandated in a top-down manner. For example, P10 explained that even the most altruistic practitioners might struggle to prioritize fairness work if it conflicts with existing incentives: *"even if everyone in their heart is thinking [fairness] is the right thing to do... they're also thinking: I have so much to do and I have all these other competing priorities... and this one is going... to get me promoted [so] how am I gonna fit in all this fairness stuff?"* (P10).

Since most companies deploying recommender systems are for-profit, many participants noted that to gain leadership buy-in, fairness evaluations had to align with financial incentives or at least not contradict them. P9 had witnessed this when setting fairness goals: *"I think how [fairness goals are decided] right now is to provide the best customer experience so the company makes the most money.... there is always like one ultimate overarching goal, which, given capitalism, is usually money."* P3 echoed this, saying that in industry, *"money is the thing that dictates everything."* P12 emphasized that in for-profit organizations, *"revenue is a priority... there always needs to be that tie to the bottom line,"* and practically, this means fairness must be tied to *"business metric[s] moving in the right direction."* P17 acknowledged that, despite criticisms, this financial focus is a practical constraint: *"if we're generating revenue, great, it means we can stick around... but if we don't generate anything... we're gone, and then none of this [fairness] work gets done [either]"* (P17).

Participants highlighted simplicity as a key factor for securing leadership buy-in for fair-

ness work, even if it meant sacrificing more impactful evaluations. This aligns with prior research suggesting that fairness metrics in industry must be simple to gain organizational support (Beattie et al., 2022). For instance, P1 said, *"for practitioners in industry, you want to just always pick the low hanging fruits, you want to pick the simplest ideas and... with the minimum effort you want to get the maximum profit basically"* (P1). P10 added that while executives can grasp technical information, *"it's easy for them to get lost in the details... so the easier it is to describe and explain a metric, the easier it is to get everyone on board with using it."* P15 also emphasized the importance of balancing appropriate fairness metrics with the ability to *"easily communicate to leadership around what does this fairness metric mean."* P18 noted that while simplicity might lead to *"worse"* fairness evaluations, it offers a practical starting point that can later evolve into more sophisticated measures, allowing for the development of better operationalizations over time.

Pragmatic Fairness

All of the constraints listed above influenced the decision-making of the participants when designing and conducting fairness evaluations, and led them towards satisficing in their work. **Satisficing** is a decision-making strategy that seeks a satisfactory or adequate outcome, rather than the optimal one (Simon, 1978). This term combines the words "satisfy" and "suffice" to describe how individuals and organizations operate within the constraints of limited information, time, and cognitive resources–and acknowledges the acceptability of imperfection. For example, P16 shared how even though some of their coworkers wanted their fairness metrics and thresholds to be *"perfect,"* in industry, there is no time nor ability to reach a perfect operationalization. This participant also described how, fortunately, in industry, *"not everything's set in stone,"* so even if an unsatisfactory fairness evaluation or intervention occurs, it can always be changed in the future to something

that is more fitting.

In addition, several participants described how in the real-world, it is not possible to holistically measure (un)fairness nor to fully solve/prevent fairness-related harms. This is due to both the subjective nature of fairness and the reductionistic assumptions we must make to measure it in practice. For example, related to evaluations, P5 said that even a theoretically *fair* system will still always *"be unfair on at least one axis,"* because, as P18 described, *"there's just too many different ways [ML systems] can be unfair."* Both P6 and P14 similarly felt that even though there is not a *"single solution,"* (P6) for how to evaluate fairness, *"at some point you just have to make some hard decisions,"* (P14) and this is a necessary part of doing fairness work in the real world. Related to *improving* fairness, P1 said, *"whatever you try to fix, it's still going to be unfair to a lot of other people... so you have to kind of choose your battles."* It has been previously argued, both theoretically and philosophically, that fairness can never truly be achieved (e.g., (Fazelpour and Lipton, 2020)). Practitioners now grapple with this reality, while still needing to make decisions and conduct evaluations to the best of their ability. Thus, fairness work in the real world often involves accepting that idealistic improvements for everyone may not be possible, but incremental progress can still be made within the constraints that shape decision-making.

I define **pragmatic fairness** as the path from constraints to satisficing in industry settings. This approach acknowledges the practical limitations and trade-offs inherent in real-world fairness evaluations and emphasizes the importance of making actionable decisions within those constraints. By satisficing, practitioners aim to address fairness concerns to a degree that is both feasible and impactful, even if the solutions are not perfect or comprehensive. Pragmatic fairness thus reflects a commitment to incremental progress, balancing competing priorities, and adapting to the dynamic and resource-limited environments in which ML systems operate. This concept highlights the need for flexibility and compromise, underscoring that striving for better–rather than perfect–fairness is essential for driving

meaningful change in industry contexts.

6.2.4 Pragmatic Fairness in Action

Based on the responses from participants, and building on frameworks mentioned in previous literature (Smith et al., 2023a; Beattie et al., 2022), in this section I explore how participants in our study conducted pragmatic fairness work (made decisions that led to satisficing as a result of the pragmatic constraints of industry). The participants conducted pragmatic fairness work through three decision-making stages of fairness evaluation: (1) scoping fairness; (2) measuring and quantifying fairness; and (3) interpreting evaluation results. I have provided Table 6.1 which outlines all of the examples that I provide in this section, based on the specific pragmatic fairness constraints and decision-making stages they align with.

Scoping Fairness Concerns

Here I outline several examples of how pragmatism shaped participants' choices about which potential fairness concerns to focus their evaluation efforts on. First, several participants described the importance of starting off with known harms and policies to get the evaluation going *quickly*, such as P2 who shared that it was more pragmatic to start with the obvious concerns instead of *"generating [a] laundry list of everything that could go wrong,"* since the evaluator will not have *"all the time in the world"* to do a *"whole assessment of all the different stakeholders that potentially could be harmed or helped with [an] application."* This method aligns with the constraint of balancing competing interests, as the company's incentives may prioritize launching the evaluation quickly, even if the ideal approach would involve taking the time to develop a more comprehensive list of fairness concerns.

Table 6.1: Examples of decisions that align with pragmatic fairness constraints throughout the stages of fairness evaluation. I discuss the idealistic counter-examples to each of these throughout Section 6.2.4.

Stage	Constraint	Examples of Pragmatic Fairness Decisions
Scope: Concerns	Competing interests	Focusing first on the known harms/risks of the system
	Getting buy-in	Deciding fairness concerns retroactively or reactively
	Lack of power/access	Internal red teaming
Scope: Stakeholders	Getting buy-in	Choosing stakeholder groups based on the marketplace
	Lack of power/access	Choosing only a few axes to analyze
Measure & Quantify	Lack of power/access	Choosing system-level metrics instead of disaggregated evaluations
	Getting buy-in	Choosing more simple metrics
Interpret	Competing interests	Optimizing fairness with respect to the business performance baseline

Another method for pragmatically determining fairness concerns was to use feedback from impacted stakeholders on social media. P16 described this kind of method as *"retroactive,"* but necessary to *"understand what type of problem you're looking for."* P15 shared they also get *"feedback either from tweets or social media or... customer service,"* and similarly described this process as *"retroactive."* This retroactive feedback approach aligns with the constraint of needing to get buy-in, because it focuses on addressing business concerns related to public relations (PR) and customer satisfaction, which could help practitioners justify fairness evaluation efforts and gain leadership support. However, I note that retroactively determining fairness concerns does require harm to occur before it can be addressed, which presents a potential adverse impact of this approach. This illustrates a tradeoff in pragmatic fairness: ideally, fairness evaluations would occur without harming impacted stakeholders. However, when practitioners are only permitted or incentivized

to evaluate fairness after harm has occurred, this approach may still offer some value.

Another common approach for participants to identify fairness concerns was to have internal team discussions and to conduct internal red-teaming to understand how a system might differentially harm certain user groups. This approach stemmed in large part from a lack of access to impacted users within the evaluation team. For example, P10 described how their team has a *"collection of test accounts with different sorts of settings... profile, country and... devices... ways to sort of simulate the experience of a broader selection of people. So we'll have, you know, a test account who is a nurse in Bangladesh or something. And follow some pages related to, you know, nurses in Bangladesh and then see what the model recommends to get a feeling for what someone in that country might really experience"* (P10). This participant mentioned that they did understand this was not the same as actually recruiting users with those profiles, but that it gave them a *"sense of what someone else's experience might be like"* in the event that they did not have access to those impacted users themselves.

Similarly, P15 shared how their team would do internal red teaming to get a sense of which *"risks that might be apparent [by] tweaking and just playing around with the model."* This participant said that their team also tries to interview internal people like engineers, product team members, and marketing team members to better assess the *"risk landscape"* of a feature (P15). P4 similarly had team discussions about potential concerns, also to make sure their model had not drifted from the original goal of the system. P16 noted that the ideal way to do this kind of internal red teaming was to *"have a diverse group of people in the room... so you're not just all thinking about it the same way."* This was a noted limitation of internal red teaming, that teams would not be able to fully understand the experience of impacted stakeholders through mock accounts, however—given the constraint of lack of access to impacted stakeholders—this method was noted to be pragmatically useful.

Scoping Stakeholders

The next step when scoping fairness evaluations is to identify which stakeholder groups are going to be involved in the evaluation. Two participants gave examples of how to make these decisions under the constraints of pragmatic fairness. P5 shared how stakeholder groups in the literature and in academia are often aligned with *"traditional social justice concerns"* like *"race, gender [or] other protected classes,"* but that in practice, the axes that they want to evaluate recommendation fairness for might look very different. In this example, P5 shared how these axes might be more related to the stakeholders of the marketplace, such as *"large versus small music producers,"* or *"genres of music,"* or *"high versus low quality songs"* in the context of a music recommendation system. This example reflects how practitioners could get greater buy-in from leadership when they align their evaluations with the organization's business priorities and financial goals, since it frames fairness work as directly relevant to the company's business context.

I note, however, that these axes may not be as unrelated to traditional social justice concerns as they initially seem. For example, the distinction between large and small music producers might reflect disparities in racial or ethnic representation, socioeconomic status, or access to resources such as marketing and distribution networks. Similarly, genres of music may be associated with specific cultural or linguistic groups, and judgments of "high" versus "low" quality songs might embed biases tied to factors like the quality of recording equipment, which can disproportionately disadvantage lower-income music producers. These intersections highlight another tradeoff of pragmatic fairness: while it would be ideal to critically examine how fairness axes are defined in practice to avoid unintentionally reinforcing broader inequities, sometimes practitioners may need to evaluate along certain axes to secure buy-in. In such cases, taking action based on what the business deems important may be more beneficial than doing nothing.

Also related to choosing evaluation axes, P15 shared how in their experience, they have developed a standardized list of *"12 or 15 different potential identity characteristics that one may have or identify as"* and that with *"unlimited time [and] resources, [they] would maybe test across all of these for every feature,"* but instead, in practice they have to think about how they can get the *"biggest bang for [their] buck with the resources [they have]"* (P15). This example highlights how practitioners are constrained by limited time and data access, which may force them to prioritize certain axes of fairness evaluation over others, even if comprehensive evaluation across all possible axes would be ideal.

Measuring and Quantifying Fairness

The next step in fairness evaluation is to measure and quantify fairness by selecting datasets, proxies, and metrics. Several participants mentioned the challenges of finding the demographic data needed to perform disaggregated evaluations in practice, and P9 and P13 shared that system-level fairness metrics were a pragmatic alternative that helped them measure unfairness in recommender systems without having to separate the evaluation by demographic groups. For example, P9 described how metrics like the Gini coefficient can help *"give you some idea about inequalities on the platform that don't necessarily tie to traditional notions of ... demographic based fairness."* This example worked within the pragmatic fairness constraint of lacking access to needed demographic data.

In the event that a practitioner *does* have access to demographic data (the ideal scenario), one participant noted that group fairness is easier to explain and communicate to their colleagues, which made it simpler to get buy-in for an evaluation: *"we find just like from a communication standpoint, group fairness is like the one that people kind of understand the best and still captures a lot of the difference[s] in user experience that we might see in the field... for the most part, we try to [choose metrics that] we can easily communicate to leadership"* (P15). In line with this constraint of choosing simple metrics to get buy-in,

P12 had also experienced this when choosing to measure fairness through the lens of bias: *"to be honest... the nice thing about [measuring fairness through bias] is it allows us to more systematically do experiments."* Even though this participant critiqued bias as a *"naive"* approach to measuring fairness, they did admit that this reduction at least *"made things possible"* (P12).

Interpreting Evaluation Results

When interpreting evaluation results, the next decision-making step is to choose a fairness *threshold* and/or *baseline*, a minimum value that must be met for the system to be deployed or shipped to end-users. One widely shared experience with pragmatic fairness evaluation between P9, P10, and P15 was the idea of omitting fairness thresholds entirely, and instead optimizing fairness within the bounds of a pre-existing business metric's threshold[1], such as P10 who shared: *"I would say let's maximize fairness, subject to a constraint of not dropping performance more than X. So I think of those performance metrics as a constraint."* P15 similarly described that they *"improve performance of fairness but still be within this like very explicit performance, like quality, threshold."* P9 added that they do not try to *"make a model with zero unfairness"* (since this is not possible), instead it is more about *"handling the rational business context of: we did what we could in order to choose the most fair model possible that still allows us to meet business performance needs"* (P9).

This fairness optimization process also involves collaborating with various teams and product owners, and P10 described how this can become an *"organizational politics thing,"* because at large companies, *"a lot of stakeholders care about a model's output, [and] there might be a whole team devoted to the needs of the person creating posts and a whole*

[1]Approximating the solution to a multicriteria optimization problem by constraining values along one dimension is a well-known technique (Ehrgott, 2005).

separate team devoted to the needs of the person viewing the post and those teams both want what's best for their people" (P10). This participant went on to describe an example where they needed to balance the needs of one team's business metrics against their own team's fairness metrics: *"we could imagine we could set a guardrail of, you know, no more than point something percent drop in some metric of feed engagement... that is the maximum amount that we can harm the viewers' experience in order to fix some fairness element of the post creators' experience"* (P10). These examples align with the constraint of balancing competing interests, since they involve negotiating between the fairness goals of the system and the performance requirements set by business teams. Perhaps in an ideal scenario, business metrics could be optimized with respect to not dropping fairness more than a certain amount. However, since business objectives are often prioritized over fairness concerns, this approach provides a pragmatic way for practitioners to manage trade-offs without compromising entirely on fairness.

6.2.5 Improving Pragmatic Fairness

In this section, I explore participants' ideas about how to improve pragmatic fairness while considering that this work inherently involves satisficing, making practical decisions that lead to imperfect outcomes. I conclude with a broader discussion about how the ML fairness research community can enhance pragmatic fairness efforts.

Including Users

One method to improve pragmatic fairness is to make more intentional efforts to include users in the process of fairness evaluation. P15 described how *"anecdotal results can be extremely [helpful]"* for motivating leadership to prioritize fairness evaluations. P9 also shared a quote from the book *Measuring Culture* (Lena et al., 2019) which says *"every*

quantitative measurement begins as a qualitative observation", and so this participant thought that *"all [quantitative] metric-based results of fairness should... inherently be supported by more qualitative observation and more qualitative work"* (P9). I note that some user involvement is more plausible than others in industry settings. For example, some companies already conduct focus groups with power users and find these methods to be useful, while conducting larger participatory co-design workshops with many impacted stakeholders might be harder and less likely pragmatically. However, recent work has shown promising methods for including users in the design of fairness evaluations, such as using focus groups to design fairness metrics with impacted stakeholders (See Study #2 in Section 4.3), incorporating responsible AI considerations into user experience (UX) efforts (Wang et al., 2023), or designing user-focused fairness evaluations to complement model/evaluation-focused ones (Deng et al., 2023).

Although approaching users might not always be an option for evaluators, sometimes the resources are there, and in these cases, several of the participants described some of the best methods for doing so. For example, P2 described how important it is to make sure the people who reach out to users have the proper skills to conduct qualitative research: *"don't send someone who can do perfect machine learning models out there to talk to some punk band about how they feel about recommendations... this is not right"* (P2). In addition, P21 shared that it is important to balance giving users enough technical information to provide useful feedback, while not overwhelming them with too many technical details of the system. *"You need to... think about to what extent to sensitize people to technical constraints and limitations. If you sensitize people to the constraints, you empower them to help design realistically implementable approaches. On the other hand, it can be challenging to do this effectively... you don't want to just give a fire hose of technical information... achieving the right balance is a delicate craft"* (P21).

Citizen Panels & Citizen Assembly

Another idea brought up by P5 and P14 was to facilitate citizen panels and assemblies as a method to make more democratic decisions about challenging fairness problems (such as which fairness metrics to use, or which fairness thresholds to set). P5 described citizen panels as the following: *"the idea here is you get a bunch of citizens together, and ideally you randomly select them... and you give them access to subject matter experts and you know lots of background material and then you have them in this small group discuss and debate [a] policy and make [a] choice... It's being... explored for exactly these types of problems where neither governments nor industry nor activists should be making the choice. What you want is a somehow representative group of regular people who are particularly well informed"* (P5). This participant also noted that both Open AI and Meta have funded citizen panels and assemblies to help them solve problems on their platforms, so this approach has precedent in real-world settings (AI, 2024; Ovadya, 2021, 2023). P14 also shared an example of how these kinds of panels could help with decisions like which fairness metrics are the most appropriate for a given use-case: *"you can imagine fairness decision-making that could look like that... you've got two or three particular fairness metrics you might use [with] significant trade-offs... [you could] spend half a day doing this and then you get them to talk it out and make a decision and that's what you go with"* (P14).

However, I note that while the idea of citizen panels evokes democratic principles, it can fall short in practice (Delgado et al., 2023). As seen with OpenAI and Meta's use of such methods, these processes may give the appearance of democracy, but they are tightly constrained, limiting public input to a narrow set of decisions (e.g., voting on model outputs or design choices). This leaves little room for citizens to contest the model's existence or critique the process itself (Young et al., 2024). Despite these limitations, citizen panels can still offer a pragmatic solution within the constraints of the system, but

I recognize that this approach may also close down opportunities for more fundamental critiques. Pragmatic fairness often involves compromises, and while such methods may provide some form of democratic participation, they may not address deeper issues or allow for the kind of substantive public debate needed for true democratic decision-making.

Evaluating Construct Validity

Another improvement to pragmatic fairness is to evaluate construct validity, when practitioners double check that their evaluations are actually measuring what they think they are measuring, so that they can trust their interpretation of the results of the evaluation. P9 described how one simple way to assess construct validity was to do a sniff-test (also introduced in prior literature (Jacobs and Wallach, 2021)) to see if their metric *"actually make[s] sense"*, and then to do empirical assessments to make sure that *"the things we've seen in the dataset that we expect to be correlated with... the target attribute are related, just to get a sense of like whether the way you're measuring your... attribute sort of aligns with like what you qualitatively think about [fairness]"* (P9). P18 also shared some examples of simple ways to assess construct validity that are similar to a sniff-test, such as writing out *"in English from the perspective of a stakeholder, what does this metric going up or down mean?"* P18 said it can also be helpful to *"think about what moves the metric.... so what different changes to the recommendations or changes to... whatever you're measuring will move the metric up or down? ... How could you move the metric without advancing your original goal? Or perhaps even undermining the original [goal]?"*. This participant described that these methods can *"be your sign you're not measuring what you were trying to measure"* or can *"lead you to the places the metric is mismatched with your goal"* (P18).

P10 also described how highly-skilled human review of fairness evaluations (such as hired contractors or domain experts) is *"the gold standard to figure out, when we're doing some-*

thing very messy and qualitative to the model, to decide whether we're doing the thing we meant to be doing" (P10). This participant further shared how these qualitative methods of human review are less pragmatic in part because they are *"expensive and slow and hard to get enough samples to really say meaningful things about [the system],"* and then went on to still say that *"despite all of those limitations, it is the only way to be really sure the model is doing what you think it's doing"* (P10).

Starting a Fairness Evaluation

Several participants shared suggestions for improving the beginning stages of fairness evaluation. P17 highlighted the benefits of involving fairness evaluators early in product development to *"progressively increase and create more robust tests as [they] know what the risk surfaces are... as opposed to just, you know, doing a one off [evaluation]"* (P17). P15 emphasized that knowing how evaluation results will be used helps structure the evaluation: *"as a team, knowing how our evaluations are going to be used also is important [for] how to structure them... even like setting the stage... at the onset, can help motivate the team towards working towards those evaluations if they know that maybe their findings should have an impact into... how the feature is developed"* (P15). P13 suggested tying fairness work to organizational success to avoid it being seen as something that *"slows down"* product development, instead presenting it as: *"if we do this, the organization will be better"* (P13).

Lastly, P21 proposed using more specific language in fairness evaluations, recognizing when the term *"fairness"* might limit the scope. They explained: *"I sometimes see almost any stakeholder need, almost anything that isn't about generating profit, described as 'fairness'. But at some point [a team I was working with] realized, no, wait, we should just say what we mean... And then they would actually tease apart and define the different aspects they've been referring to as 'fairness', [and] then actually thinking about how to*

address these particular issues. So I have seen a big difference between teams that use 'fairness' in a scoped way versus teams where it somehow ends up becoming everything [besides] business needs" (P21). The critique that the term "fairness" might constrain our ability to effectively evaluate a system for harm is a significant consideration—one that I revisit in Section 7.1.

6.3 Conclusion & Pathways Forward

Pragmatic fairness parallels the concept of "muddling through," introduced by Lindblom (2018), which captures the reality of decision-making as a continuous, imperfect process prioritizing practicality and adaptability over unattainable idealized rationality. While this approach to fairness is open to criticism—particularly because it aims to satisfice within business constraints rather than strive for ideal solutions—its value lies in addressing real-world complexities. Fazelpour and Lipton (2020) argue that many existing fairness metrics are grounded in ideal theory, which assumes a perfectly just world and often neglects the nuanced challenges of societal application. Such idealized frameworks risk producing interventions that, in practice, are ineffective or even harmful. To counter this, the authors advocate for a non-ideal theoretical perspective that accounts for societal injustices and practical constraints. I position pragmatic fairness as an empirical extension of these philosophical arguments, offering a complementary lens to operationalize fairness in real-world contexts.

I acknowledge that pragmatic fairness inherently involves trade-offs. Simplifying proxies and metrics, creating mock user accounts, and treating fairness as a reactive rather than preventative process are all examples of pragmatic approaches that risk exacerbating harm to impacted stakeholders. Nevertheless, I argue that there is space within the ML fairness research community for both idealistic and pragmatic visions of ML fairness, both *nyaya* and *niti*, in Sen's terminology (Sen, 2009). Not all work in this field can focus on critiquing and dismantling real-world systems, nor can it overlook the harms that arise from operating within them. It is important to strike a balance between working within the system and working to fundamentally change the system.

I propose that pragmatic fairness can serve as a foundation for inspiring future work aimed at improving fairness within the constraints of industry. By highlighting constraints and

providing tangible examples, I hope to help ML fairness researchers align their work with the real-world motivations, limitations, and power dynamics that shape its application. For instance, future research could enhance internal red-teaming efforts to better capture the perspectives and concerns of impacted stakeholders, particularly in cases where practitioners lack direct access to these groups. Additionally, there is a pressing need for educational resources on ML fairness tailored to organizational leadership, reducing the burden on practitioners to secure buy-in. Another valuable contribution could be to develop guidance to help practitioners prioritize fairness concerns while balancing business objectives, ensuring that fairness efforts align with, rather than hinder, product and organizational goals.

While achieving perfection within existing constraints may be unrealistic, pursuing marginal improvements is both meaningful and essential. I hope the examples and insights provided in this chapter serve as starting points for refining pragmatic fairness practices, while recognizing that broader advancements in fairness evaluation extend beyond the scope of this work and merit further exploration.

Chapter 7

Bigger Picture

> **"Measure what is measurable, and make measurable what is not so."**
>
> – Galileo Galilei

This dissertation has examined the challenges and opportunities of incorporating fairness into the design of multistakeholder recommender systems and machine learning more broadly. Based on insights from users, practitioners, and ML fairness experts, I have explored potential best practices for fairness operationalization in real-world, multistakeholder recommender systems that takes into consideration the varied needs of an organization and its stakeholders.

In this dissertation, I have also introduced the concept of pragmatic fairness, which outlines the path from business-driven constraints to making satisfactory, though imperfect, decisions when evaluating ML fairness in industry settings. Through this lens, I have explored how fairness evaluation might successfully balance theoretical ideals with industry realities, highlighting tradeoffs and pointing toward future work that aims to bridge this gap. This conclusion chapter synthesizes my key insights from all five studies in this dissertation, and offers a roadmap forward for researchers and practitioners working at

the intersection of fairness and machine learning.

7.1 Takeaways & Implications

In this section, I synthesize the major takeaways and implications of this dissertation, offering insights into the practical and conceptual lessons learned through my research and their broader significance for the field of responsible AI.

Fairness Is Not One-Size-Fits-All

First, I have learned and relearned that fairness will never be a one-size-fits-all construct. There will never be a universal method to measure fairness that satisfies every individual and stakeholder within a system. Individual users often have different fairness preferences, as do groups of impacted stakeholders, practitioners balancing competing business needs, and the organization at large, which may prioritize revenue, regulatory compliance, and public perception over all else. I have seen these differing needs and perspectives towards fairness firsthand throughout all of the studies in this dissertation. Although I gathered perspectives from 102 participants in total for this dissertation, each participant brought unique experiences, preferences, and conceptualizations of fairness, with no two individuals sharing exactly the same perspective.

This variability of fairness preferences also makes it incredibly challenging for practitioners to decide whose preferences to prioritize. For example, in Study #3, I observed that certain fairness logics may be inherently incompatible. For instance, choosing to rank high-risk loans higher on the Kiva platform could benefit high-risk borrowers and generate positive outcomes for individuals living in volatile commodity markets. However, this approach also increases the risk of loan defaults for lenders, potentially reducing the funds

available for future lending within the online marketplace. In contrast, ranking low-risk loans higher is better for lenders, but further marginalizes and excludes borrowers who live in high-risk areas. This complexity highlights how different fairness implementations can prioritize certain stakeholders over others, often leaving practitioners without a clear "best" or "optimal" solution.

I also discussed this lack of an "optimal" solution in Section 2.2.3. As Selbst et al. (2019) note, *"there is no way to arbitrate between irreconcilably conflicting definitions [of fairness] using purely mathematical means."* This complexity is compounded by measurement challenges, as different user groups interact with systems in diverse ways, often skewing metrics despite efforts to account for such differences (Mehrotra et al., 2017; Stray et al., 2022). For instance, Stray et al. (2022) demonstrate how metrics like dwell time can systematically misrepresent certain demographics, such as older versus younger users. This means that even measurement can be a fairness issue itself. These challenges reveal that fairness cannot always be perfectly defined, measured, or optimized, yet **practitioners must still make decisions about how to operationalize it.**

> "I think it's hard, generally, [to decide] who deserves fairness more than the others. The answer is... everybody at the same time. But there needs to be a decision. **Somebody has to make a decision**" (P1, Study #5).

This presents both opportunities and challenges for operationalizing fairness. On one hand, we have a diverse array of approaches to evaluating and embedding fairness into systems, where no single method is inherently correct or incorrect but rather more or less suited to the specific context. On the other hand, it highlights the unsolvable nature of this work–there will never be a definitive solution to the problem of fairness, and every approach will involve trade-offs that impact stakeholders in different ways.

This is why I believe there is a critical need for more guidance on scoping fairness eval-

uations to identify which aspects of fairness are most relevant for a given context and how to prioritize competing ideals. This was especially apparent after conducting Study #4, where many practitioners who I spoke to shared how they would struggle to decide between various ML fairness evaluation approaches in large part because they all seemed equally important and applicable to their evaluation goals. This was also apparent in Study #3, where many fairness logics might be equally appropriate to enact on a platform–all of which might differentially prioritize certain stakeholders. Practitioners would benefit from guidance on how to choose which fairness logic to enact in a given context.

I envision this guidance taking the form of workshops or facilitated discussions that bring together impacted stakeholders, domain experts, modeling teams, and ML fairness evaluators. These sessions could help everyone collaboratively identify all potential viable approaches, prioritize them based on the group's diverse and often competing priorities, and provide structured support in justifying the chosen approach. This and other kinds of scoping guidance would help practitioners navigate the inevitable trade-offs between optimizing fairness for some while acknowledging that not all needs can be simultaneously met.

Users' Perspectives on Fairness

Another important implication of the work in this dissertation is the role that impacted stakeholders and users could and should play in fairness evaluation design.

I see two distinct approaches for including users in the design of fairness evaluations: (1) to gather feedback on the overall design of fairness operationalization; or (2) to understand users' specific lived experiences of unfairness. The first approach could be pragmatically accomplished through participatory methods such as focus groups with power users,

surveys, or democratic methods as discussed in Section 6.2.5. Although many of these methods may currently fall short in practice, they pose promising directions for future work to pragmatically include stakeholders in fairness evaluations going forward.

While the second approach (understanding users' experiences of unfairness) shows significant promise–as evidenced by Study #2 of this dissertation–it comes with a notable challenge: users' perceptions of fairness often reflect their overall satisfaction with the system rather than an objective sense of what they "deserve." For instance, in Study #2, content creators frequently described feeling they were treated unfairly when they did not achieve the views, engagement, or exposure they desired. This phenomenon, which I term "selfish conceptions of fairness," creates a complex evaluation landscape. Systems designed to meet individualized expectations of fairness would face inherent limitations, as it is virtually impossible to satisfy every user's ideal outcome. Future research would benefit from exploring ways to integrate this individualized perspective of fairness into system design while balancing the reality that users often prioritize their own interests.

During this dissertation, I also identified two distinct enactments of fairness: (1) perceptions of fairness, grounded in the lived experiences of impacted stakeholders, where fairness interventions are designed to be perceptible and meaningful to users; and (2) theoretical notions of fairness, aimed at improving systems more holistically through high-level fairness considerations that may not be immediately perceptible to users but seek to create long-term, systemic benefits. There are a variety of research studies that have focused on capturing peoples' perceptions of ML fairness (e.g., (Woodruff et al., 2018; Alkhathlan et al., 2024; Kingsley et al., 2022; Harris et al., 2023; Kasinidou et al., 2021; Binns et al., 2018)), and similarly there have been many studies that have focused on capturing theoretical notions of ML fairness (e.g, (Hardt et al., 2016; Berk et al., 2021; Agarwal et al., 2018; Dwork et al., 2012; Buolamwini and Gebru, 2018; Zafar et al., 2017; Kleinberg et al., 2016; Grimmelmann, 2010; Zehlike et al., 2017; Singh and Joachims, 2018) just to name

a few), and one study also conducted a fascinating attempt to combine both theoretical notions and perceptions of ML fairness (Srivastava et al., 2019).

I believe the ML fairness research community would benefit from more open discussions about these different conceptualizations of fairness. Without clearer delineations between fairness grounded in opportunities, outcomes, processes, or perceptions–efforts to operationalize fairness risk becoming more complex than they already are. Moreover, a deeper understanding of the trade-offs and consequences of prioritizing one approach over another is crucial. By anchoring research within one or more of these frameworks, we can better refine and strengthen our approaches to fairness operationalization.

Idealistic vs. Pragmatic Fairness

The distinction between idealistic fairness and pragmatic fairness stands out as a key takeaway in this dissertation, especially as highlighted in Study #5, and previously discussed in Section 6.2.5. These two approaches–aspiring to ideal fairness versus implementing practical (though imperfect) solutions–are not mutually exclusive but complementary.

Idealistic fairness pushes the boundaries of what is possible, challenging systems to strive for fairness that is theoretically perfect, while pragmatic fairness recognizes the real-world constraints and imperfections that shape the operationalization of fairness in complex settings. This dual approach is crucial because fairness is inherently contextual, and there is no one-size-fits-all solution that can address every concern from all stakeholders. Ideal fairness offers the vision and aspiration necessary for long-term progress, while pragmatic fairness ensures that systems can actually function in practice and address pressing inequalities in a way that is actionable. This dual focus is vital because it allows fairness to be both aspirational and achievable, ensuring that systems are not only ethically designed but also practically viable.

Throughout this dissertation, several participants also emphasized that idealistic assumptions about fairness can sometimes hinder effective operationalization. For example, in Study #2, one participant suggested that dating app recommendations would be fairer if they promoted greater diversity, helping to address issues like sexual racism. However, they pointed out that this idealistic vision assumes all users are kind, respectful, and accepting—an assumption that does not align with reality. Implementing such a system could inadvertently lead to harmful outcomes, such as hateful messages or interactions when users are shown diverse profiles they consider unattractive, "not in their league," or misaligned with their prejudices toward certain identity demographics. A similar sentiment emerged in Study #5, where a participant noted that fairness operationalization for dating apps must account for the inherent popularity of certain profiles—an unavoidable reflection of the broader inequities in our society. These are both great examples of how focusing on idealistic solutions when operationalizing fairness might lead to unintended consequences, and that there is utility in recognizing the practical realities of unfairness in society when designing fairness evaluations and interventions.

That said, I believe there is an opportunity to address societal unfairness through thoughtful technology design—or at the very least, to avoid exacerbating it. Operating within an inherently unfair system does not mean we must perpetuate its inequities. With careful consideration of the context in which recommender systems operate, we can identify ways to gradually mitigate some of this unfairness. For instance, as demonstrated in Study #3, fairness can be promoted within a recommendation ecosystem through UI design, such as by incorporating carousels that highlight items from underrepresented provider groups. Similarly, in Study #1, some participants suggested embedding an organization's fairness goals directly into the website to raise user awareness of provider-side fairness issues and to encourage actions that align with these goals. There are numerous strategies for addressing fairness, discrimination, or harm on a platform that extend beyond datasets, recommendation models, or re-ranking algorithms. These UI-focused approaches hold

significant promise, and I encourage future research to explore fairness operationalization through interface design.

In this dissertation, I argue that the ML fairness research community must strike a balance between idealistic and pragmatic fairness to ensure that our efforts are both meaningful and impactful. Operationalizing fairness within business constraints alone runs the risk of stifling innovation and limiting the potential for broader societal change. If fairness research focuses too heavily on pragmatic solutions, it may inadvertently reinforce existing power structures and fail to address deeper systemic issues that require idealistic aspirations and transformative goals. On the other hand, prioritizing idealistic fairness without accounting for the real-world limitations of practitioners could result in a lack of tangible progress. When fairness ideals are disconnected from practical implementation, there is a risk of creating frameworks that are theoretically appealing but difficult to translate into meaningful action.

While I am a strong advocate for idealistic philosophical exploration as a purely intellectual pursuit—I believe it is essential to understand fairness at a deep, conceptual level—I maintain that integrating both idealistic and pragmatic approaches is essential for the future of fairness research in ML. This balance can encourage a dynamic dialogue between theory and practice, where idealistic concepts can inform and inspire pragmatic solutions, and vice versa.

In line with this approach, I also believe practitioners could benefit from speculative design practices (Dunne and Raby, 2024), which involve imagining alternative futures and challenging conventional thinking. Speculative design, as an approach, encourages us to envision possible futures in which fairness is realized in unexpected ways, pushing us to think beyond the constraints of current technologies and societal norms. By integrating speculative design into fairness evaluation design, practitioners can explore the boundaries of what is possible and consider how fairness might be realized even within industry

constraints. This method encourages creative, transformative thinking that can inspire innovative solutions to fairness challenges, rather than being limited by existing paradigms. Additionally, by working through speculative scenarios, practitioners can engage with diverse stakeholder groups to anticipate the potential impacts of different fairness approaches, thereby strengthening the robustness of fairness evaluations and ensuring that they are forward-thinking and adaptable to future technological and social shifts.

Together, the combination of idealistic fairness, pragmatic considerations, and speculative design practices offers a promising path forward for the ML fairness community, one that can generate both theoretical depth and practical impact.

The Limitations of Fairness

Finally, I want to highlight several limitations associated with focusing exclusively on fairness as the primary lens for responsible AI. While fairness is undoubtedly a crucial aspect of AI systems, it is only one of many beneficial properties that should be considered when evaluating the broader ethical implications of these technologies. Other factors, such as mitigating additional representational and allocative harms, improving transparency and explainability, and addressing the negative societal impacts of AI systems, are equally important. These concerns go beyond fairness, as they address the broader and more complex landscape of responsible AI.

For example, transparency and explainability are essential to fostering user trust and enabling accountability, but they do not always directly relate to fairness (Doshi-Velez et al., 2017). I do note that in this dissertation work, there did appear to be an important connection between transparency, explainability, and perceptions of fairness from users. For instance, during Study #2, a participant from the dating app focus group expressed that greater transparency about how the recommendation algorithm works would make

the system *feel* more fair to them, even if the system itself was not theoretically fair in practice. This sentiment was echoed by participants in Study #1, who indicated that transparency regarding fairness goals was important to them, even if they were not directly involved in defining those goals. These findings suggest that addressing other responsible AI considerations, such as transparency and explainability, could enhance user perceptions of fairness, even when improving fairness itself may not be the primary objective.

I also learned during this dissertation work that using the word "fairness" when conducting resonsible AI work can be a limitation itself. This was showcased during Study #5, when one participant pointed out that using the lens of "fairness" to evaluate a system for harm could limit the scope of the evaluation in a negative way. This also arose during Study #4, when several practitioners from Spotify noted that the word "fairness" was confusing for them, and they thought it would be eaiser to just frame their evaluation based on the exact criteria they were trying to capture. These perspectives highlight the need for a more holistic view of responsible AI operationalization that extends beyond fairness. Focusing solely on fairness may result in overlooking other critical harms, such as those related to safety, privacy, and autonomy, which can be just as detrimental, if not more so, in specific contexts (O'Neil, 2017).

The reductionist framing of fairness as a main ethical concern could also lead to the reification of fairness metrics that fail to capture the lived experiences of individuals who are most affected by AI systems. For example, during Study #2, I observed that several focus group participants developed "fairness metrics" that reflected their lived experiences of unfairness, yet these metrics did not align with traditional or theoretical conceptions of fairness. This showcases the gap between how we perceive fairness or unfairness in our everyday lives and theoretical definitions of fairness, suggesting that our efforts to quantify fairness may not always capture the full complexity of these perceptions. Consequently, it is crucial for the machine learning fairness research community to critically evaluate the

benefits and limitations of framing certain ethical concerns through the lens of fairness.

Future work might consider how decoupling fairness evaluations from theoretical and normative frameworks could provide a more nuanced and context-specific approach to responsible AI. By shifting towards a broader framework that incorporates a wider range of ethical concerns, researchers could develop more comprehensive evaluations that better capture the real-world impacts of AI technologies. For instance, integrating concepts from justice, ethics of care, and human-centered design could help guide the development of AI systems that prioritize the well-being of all stakeholders involved. Embracing this shift would allow us to move towards creating AI technologies that augment human capabilities and foster social good, rather than exploiting vulnerable populations or reinforcing existing power imbalances.

7.2 Where Do We Go From Here?

As fairness in machine learning continues to evolve, numerous open questions remain about how to best operationalize fairness, incorporate stakeholder perspectives, and balance competing values in real-world systems. Below, I outline some of the most pressing challenges in the form of questions, and propose potential pathways forward.

How can we reconcile differences between perceptions of fairness and theoretical notions of fairness?

One key challenge that arose from this dissertation is how to bridge the gap between subjective experiences of fairness and formalized fairness metrics. A promising avenue for future work could be to translate end-user perceptions into quantitative measures. For instance, the fairness metrics designed by providers in Study #2 could be refined in

collaboration with data scientists, statisticians, and other domain experts to transform qualitative insights into actionable system-level fairness measures. These metrics could then be tested empirically to assess their feasibility.

Alternatively, instead of focusing on individual-level perceptions of fairness, we could design studies where users engage with existing fairness frameworks and provide structured feedback on their applicability. By shifting the conversation toward system-wide fairness concerns, we may be able to incorporate user needs without being constrained by the inherently subjective—and sometimes self-serving—nature of individual fairness perceptions.

What would it look like to incorporate users' fairness needs into ML design?

Something I have learned from this dissertation work is that addressing users' fairness concerns in recommender systems may not always require algorithmic interventions. Instead, UI/UX and platform design changes could play a crucial role. For example, platforms could introduce configurable fairness settings, allowing end-users to adjust how much fairness is incorporated into their recommendations. Accompanying this with educational resources would help users understand fairness goals and their impact on personalization.

For providers, greater transparency and explainability are crucial. Enhancing visibility into how and why content is surfaced—or not—could empower providers to make informed decisions about how they interact with the recommendation ecosystem. Some providers may prioritize accurate exposure over sheer reach, so offering tools to understand audience demographics could better align recommendation outcomes with provider needs. Additionally, platforms could implement feedback mechanisms that allow end-users and providers to report perceived unfairness. This could include mechanisms for user feedback to understand: (1) when users feel they are being treated unfairly; (2) why they feel this

way; and (3) what they believe they would need to correct this unfairness. This would also be an opportunity to learn about the experiences and needs of users without having to overcome the obstacles that often prevent user studies from happening in industry.

How can we disentangle user perceptions of fairness from user satisfaction? Should we?

Conflating fairness with user satisfaction risks allowing fairness concerns to be driven by individual desires rather than broader ethical considerations. While fairness and satisfaction may sometimes align, they serve distinct purposes. If fairness is to be conceptualized as more than just user preference optimization, then we must develop strategies to disentangle the two.

One potential approach is educational: for example, content providers competing in a zero-sum recommendation environment could be made aware of the trade-offs involved. If a provider believes they deserve higher ranking, they should also recognize that this would mean demoting another provider. Encouraging such reflection may help separate fairness from mere self-interest and guide fairness evaluations that consider broader stakeholder impacts.

Should all technical ML practitioners be fairness experts?

Not all ML practitioners need to be fairness experts, but some must be—and we need a more structured division of expertise. Ideally, fairness work should involve three types of expertise: **social science and humanities scholars** who understand ethics, value tensions, and qualitative aspects of fairness; **technical experts** who deeply understand ML models, data pipelines, and algorithmic capabilities; and **interdisciplinary practitioners** who can bridge these two groups, facilitating communication and ensuring fairness concerns are technically and socially informed.

In practice, many fairness experts emerge organically within organizations rather than through formal training. However, structured pathways for developing fairness expertise—through coursework, mentorship, and role specialization—could strengthen fairness efforts in industry.

What should be done to support ML practitioners in applying fairness pragmatically?

One of the biggest challenges for fairness operationalization is the lack of structured guidance to help ML practitioners navigate complex trade-offs in real-world systems. While there has been significant academic work on fairness metrics and interventions, practitioners often struggle to apply these insights due to misalignment with business priorities, data constraints, and unclear organizational incentives.

A common approach to addressing this gap has been the development of open-source fairness toolkits, but these resources have limitations. In practice, fairness toolkits often provide guidance that is too abstract or high-level to be actionable because the creators lack access to the proprietary data, system constraints, and business objectives that shape real-world ML development. Without this contextual knowledge, toolkits risk offering recommendations that are impractical or disconnected from the realities of deployment. Future efforts could explore more context-aware, industry-specific guidance that acknowledges these constraints and provides more tailored pathways for implementation.

Another promising direction is increasing transparency within industry about fairness efforts. One participant from Study #5 shared with me that they have seen practitioners in companies with strong internal support for fairness work consider publicly sharing details about their initiatives. If organizations with robust fairness programs become more open about their work—whether through publications, conference talks, or corporate social responsibility reports—this could establish new industry norms. By making fairness

work more visible, organizations that neglect fairness considerations may face reputational risks, potentially creating a competitive incentive to engage with fairness proactively. If fairness becomes an expectation rather than an optional luxury, companies that fail to invest in it may be perceived as lagging behind.

Finally, a crucial but often overlooked factor is how fairness education is marketed to industry leadership. Many companies deprioritize fairness work because it is seen as a compliance issue rather than a business imperative; this mimics institutional barriers that computing educators face when incorporating ethics into their classrooms because of ethics being seen as unrelated to computing curricula (Smith et al., 2023d). However, fairness interventions can improve brand image, prevent negative PR, and increase user trust, satisfaction, and retention—all of which directly contribute to revenue. If fairness education and workshops are framed as strategic assets that reduce regulatory risk, improve public perception, and drive long-term user engagement, leadership may be more willing to invest in practitioner education. Future efforts should explore how to effectively communicate the business case for fairness, ensuring that practitioners receive the institutional support needed to incorporate fairness into ML development in meaningful ways.

How might focusing on fairness in process, rather than just fairness in outcomes, improve ML fairness work?

Many fairness concerns raised by the providers in Study #2 related not just to fairness of outcomes but also to the fairness of the underlying decision-making process. Yet, because users typically only see outcomes, when we ask users' about the fairness of a system, they often rely solely on subjective assessments of these outcomes.

Future research could explore how increasing transparency about system processes impacts fairness perceptions. For example, instead of asking users to describe their perceptions of

fairness with respect to the *outcomes* they experience on the platform, we could present users with descriptions of fairness *processes* and ask them to share their perspectives on whether they believe those processes to be fair or not. This could provide a way to assess fairness at both the individual and system levels while also reducing reliance on subjective outcome-based fairness assessments.

Another related question that has come up throughout this dissertation is *whose needs should be prioritized when operationalizing fairness?* The answer to this question is not clear nor simple; prior research has shown that practitioners, users, and the general public have different perspectives about what they value with respect to responsible AI efforts (Jakesch et al., 2022). Part of improving fairness in ML processes might also include being more transparent and explicit about whose values and needs are prioritized and why.

How can we balance working within the system versus working to fundamentally change that system?

At the 2023 FAccT conference, the "Theories of Change" CRAFT session (Wilkinson et al., 2023) revealed deep divisions within the fairness research community regarding whether responsible AI ought to be done through reform versus abolition. Many researchers in this discipline advocate for dismantling biased systems entirely, while many others work within existing institutions to mitigate harm.

Both perspectives are necessary. Activists and policymakers who push for systemic change help our discipline ensure that larger ethical shifts occur, while practitioners who are embedded in industry play a critical role in making immediate improvements. If only the former exists, while we wait for larger shifts to occur, industry may cause more harm that goes unchecked. If only the latter exists, incremental change may occur, but the system as a whole may remain fundamentally flawed. I believe a balance must be struck. In

order to effectively conduct fair ML work, we must foster collaboration between these two ideologies in our research community rather than viewing them as mutually exclusive.

7.3 Summary

There is no doubt that incorporating fairness into machine learning systems is an important consideration in the design of these technologies. The process of operationalizing the value of fairness into a technical system requires translating theoretical, unobservable, and contested concepts into empirically observable proxies. However, this translation process is not straightforward, and if done poorly, runs the risk of causing more harm rather than alleviating it.

In the domain of recommendation, operationalizing fairness is a relatively new endeavor that lacks critical guidance for practitioners. This lack of guidance coupled with the proliferation of available fairness metrics and the business-minded constraints of industry has led to a complex decision-making space that can leave practitioners unsure where to even begin conducting fairness work.

Through five qualitative research studies with users, practitioners, and ML fairness experts, I have explored and reported the challenges, needs, desires, considerations, and implications of these perspectives on the design of ML fairness operationalization for real-world recommender systems.

Through this work, I gained several key insights. First, I learned that end-users and providers of recommendation systems want to be included in fairness design, and their contributions can be invaluable to the process. Second, practitioners need clearer guidance when scoping fairness evaluations, particularly when navigating multiple equally valid approaches. Finally, a balance between idealistic aspirations and pragmatic constraints is

essential for effective fairness operationalization. By embracing these lessons, I believe we can create responsible AI systems that are better aligned with the diverse needs of their stakeholders and the broader society they aim to serve.

7.4 Concluding Remarks

I was first introduced to machine learning fairness seven years ago through Hardt et al. (2016)'s groundbreaking paper on measuring equality of opportunity in supervised learning[1]. This influential work sparked my initial exploration into fairness metrics, particularly from a quantitative perspective. I was captivated by the idea that a philosophical concept could be captured through quantitative evaluation. Seven years and one dissertation later, my belief in the value of this approach remains steadfast.

However, I've come to realize that numbers alone are insufficient. While they enable large-scale evaluations and allow for optimization of systems based on very specific criteria, they fall short of capturing the complex, lived experiences and unique needs of real people and their communities. This is where stories become invaluable. This dissertation is a tribute to those stories: the diverse perspectives, experiences, and visions of the individuals deeply involved in the operationalization of ML fairness.

I've designed, discussed, and collaborated with over one hundred research participants: ML fairness experts, practitioners, and users of recommender systems—and they have taught me what machine learning fairness is, and what it *can be*.

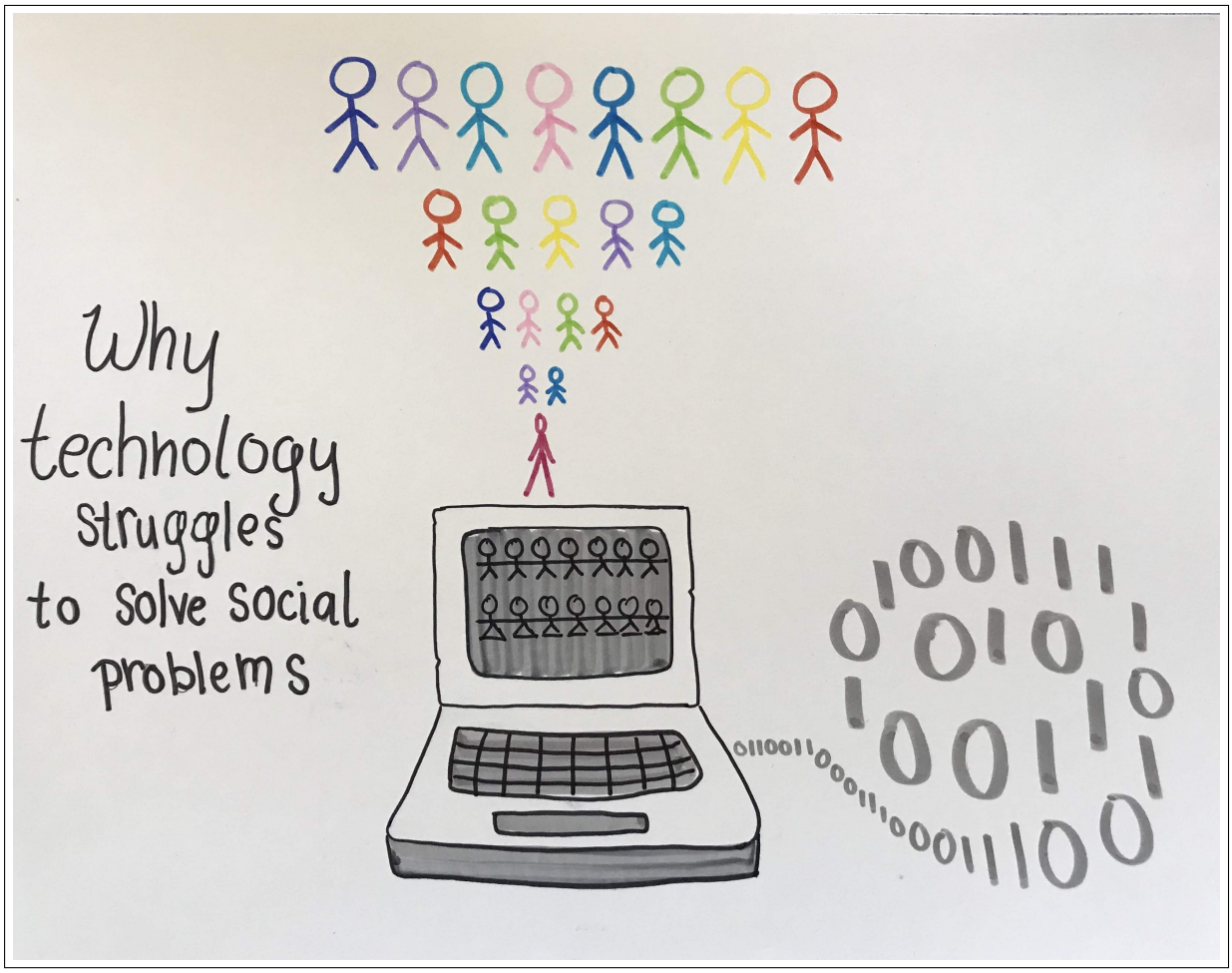
I have now reached the point where I bring this dissertation to a close. I truly hope this

[1]During this time, I also created a tutorial that provides an introduction to the basics of quantifying ML fairness that is based on this paper by Hardt et al. (2016), it was intended to be used in introductory data science courses, and is still available for use at github.com/jesmith14/fairness-ml-tutorial.

work is useful to my research community, to practitioners, and ideally—to those who are most impacted by technology.

I do not believe that we will ever be able to actually measure fairness, not in a holistic way at least. However, I do believe that if we can foster mindful collaborations between users, practitioners, and experts while balancing idealistic and practical approaches to ML fairness, we can do our best to keep trying.

Fin.



(I drew this picture in 2019, during the first year of my PhD program.)

Bibliography

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the Thirty-Second International FLAIRS Conference*. AAAI Press, Washington, DC, USA, 413–418.

Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

Open AI. 2024. Democratic inputs to AI grant program: lessons learned and implementation plans. https://openai.com/index/democratic-inputs-to-ai-grant-program-update/. [Accessed 22-12-2024].

Ken Alder. 2003. The measure of all things: The seven-year odyssey and hidden error that transformed the world.

Sanna J Ali, Angèle Christin, Andrew Smart, and Riitta Katila. 2023. Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 217–226.

Mallak Alkhathlan, Kathleen Cachel, Hilson Shrestha, Lane Harrison, and Elke Rundensteiner. 2024. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1940–1970.

Lameck Mbangula Amugongo, Nicola J Bidwell, and Caitlin C Corrigan. 2023. Invigorating Ubuntu Ethics in AI for healthcare: Enabling equitable care. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 583–592.

Kenton Anderson and Gregory D Saxton. 2016. Smiles, babies, and status symbols: The persuasive effects of images in small-entrepreneur crowdfunding requests. *International Journal of Communication* 10 (2016), 1764–1785.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 249–260.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

Julia Angwin and Terry Jr. Parris. 2016. Facebook lets advertisers exclude users by race. `https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race`

Richard Arneson. 2002. Equality of opportunity. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Palo Alto, CA, USA. `https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/`

Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 24–35.

Solon Barocas, Elizabeth Bradley, Vasant Honavar, and Foster Provost. 2017. Big data, data science, and civil rights. arXiv preprint arXiv:1706.03102.

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 368–378.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and Machine Learning. fairmlbook. org, 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

Eric PS Baumer and M Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2271–2274.

Lex Beattie, Dan Taber, and Henriette Cramer. 2022. Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 528–530.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

Nicole Bergen and Ronald Labonté. 2020. "Everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qualitative health research* 30, 5 (2020), 783–792.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

Marya L Besharov and Wendy K Smith. 2014. Multiple institutional logics in organizations: Explaining their varied nature and implications. *Academy of Management Review* 39, 3 (July 2014), 364–381. `https://doi.org/10.5465/amr.2011.0431`

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, and Ed H Chi. 2019a. Fairness in recommendation ranking through pairwise comparisons. arXiv:1903.00780 [cs.CY]

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019b. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. `https://doi.org/10.1145/3292500.3330745`

Amy Binder. 2007. For love and money: Organizations' creative responses to multiple environmental logics. *Theory and Society* 36, 6 (Dec. 2007), 547–571. `https://doi.org/10.1007/s11186-007-9045-x`

Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 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*. 1–14.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 100205.

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The forgotten margins of AI ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 948–958.

Linda Birt, Suzanne Scott, Debbie Cavers, Christine Campbell, and Fiona Walter. 2016. Member checking: a tool to enhance trustworthiness or merely a nod to validation? *Qualitative health research* 26, 13 (2016), 1802–1811.

Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.

Amanda Bower, Kristian Lum, Tomo Lazovich, Kyra Yee, and Luca Belli. 2022. Random Isn't Always Fair: Candidate Set Imbalance and Exposure Inequality in Recommender Systems. (2022). `https://doi.org/10.48550/ARXIV.2209.05000`

Janet M Box-Steffensmeier, Henry E Brady, and David Collier. 2008. *The Oxford handbook of political methodology*. Vol. 10. Oxford Handbooks of Political.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

Virginia Braun and Victoria Clarke. 2021a. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

Virginia Braun and Victoria Clarke. 2021b. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative research in sport, exercise and health* 13, 2 (2021), 201–216.

Meredith Broussard. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press.

Richard Buchanan. 2001. Human dignity and human rights: Thoughts on the principles of human-centered design. *Design issues* 17, 3 (2001), 35–39.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Proceedings of Machine Learning Research, Vol. 81. PMLR, New York, NY, USA, 77–91. `https://proceedings.mlr.press/v81/buolamwini18a.html`

Robin Burke. 2017a. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

Robin Burke. 2017b. Multisided fairness for recommendation. 2017 Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017). arXiv:1707.00093 [cs.CY]

Robin Burke, Nicholas Mattei, Vladislav Grozin, Amy Voida, and Nasim Sonboli. 2022a. Multi-agent social choice for dynamic fairness-aware recommendation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*. ACM Press, New York, NY, USA, 234–244. https://doi.org/10.1145/3511047.3538032

Robin Burke, Pradeep Ragothaman, Nicholas Mattei, Brian Kimmig, Amy Voida, Nasim Sonboli, Anushka Kathait, and Melissa Fabros. 2022b. A performance-preserving fairness intervention for adaptive microfinance recommendation. In *Proceedings of the KDD Workshop on Online and Adapting Recommender Systems (OARS)*. ACM Press, New York, NY, USA, 6 pages.

Robin Burke, Amy Voida, Nicholas Mattei, and Nasim Sonboli. 2020. Algorithmic fairness, institutional logics, and social choice. In *Harvard CRCS Workshop: AI for Social Good*.

David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

Carlos Castillo. 2019. Fairness and transparency in ranking. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 64–71.

Sushma Channamsetty and Michael D Ekstrand. 2017. Recommender response to diversity and popularity bias in user profiles. In *Proceedings of the Thirtieth International FLAIRS Conference*. AAAI Press, Washington, DC, USA, 657–660.

Jaegul Choo, Changhyun Lee, Daniel Lee, Hongyuan Zha, and Haesun Park. 2014. Understanding and Promoting Micro-Finance Activities in Kiva.Org. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 583–592. https://doi.org/10.1145/2556195.2556253

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A Feder Cooper, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, et al. 2024. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. *arXiv preprint arXiv:2412.01934* (2024).

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

Civil Rights Act. 1964. Civil Rights Act of 1964. (1964). https://www.govinfo.gov/app/details/COMPS-342

Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018a. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018b. Assessing and Addressing Algorithmic Bias in Practice. *Interactions* 25, 6 (oct 2018), 58–63. https://doi.org/10.1145/3278156

Henriette Cramer, Jenn Wortman Vaughan, Ken Holstein, Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, and Sravana Reddy. 2019. Challenges of incorporating algorithmic fairness into industry practice. FAT* 2019 Tutorial. https://www.microsoft.com/en-us/research/uploads/prod/2020/10/FAT_2019-tutorial_-algorithmic-bias-in-practice.pdf

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922* (2022).

Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

Michael Ann DeVito. 2022. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.

Michael Ann DeVito, Ashley Marie Walker, and Julia R Fernandez. 2021. Values (mis) alignment: Exploring tensions between platform and LGBTQ+ community design values. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

Karlijn Dinnissen and Christine Bauer. 2023. Amplifying Artists' Voices: Item Provider Perspectives on Influence and Fairness of Music Streaming Platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 238–249.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

Anthony Dunne and Fiona Raby. 2024. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming*. MIT press.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Ben Eggleston. 2014. Act utilitarianism. In *The Cambridge Companion to Utilitarianism*, Ben Eggleston & Dale E. Miller (Ed.). Cambridge University Press, Cambridge, UK, 125–145.

Matthias Ehrgott. 2005. *Multicriteria optimization*. Vol. 491. Springer Science & Business Media.

Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, New York, NY, USA, 172–186. https://proceedings.mlr.press/v81/ekstrand18a.html

Mehdi Elahi, Himan Abdollahpouri, Masoud Mansoury, and Helma Torkamaan. 2021. Beyond algorithmic fairness in recommender systems. In *Adjunct proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 41–46.

Lydia Emmanouilidou. 2018. https://www.pri.org/stories/2018-02-16/what-can-ai-learn-non-western-philosophies..

Adalbert Evers. 2005. Mixed welfare systems and hybrid organizations: Changes in the governance and provision of social services. *International Journal of Public Administration* 28, 9–10 (2005), 737–748. https://doi.org/10.1081/PAD-200067318

Jaydon Farao, Ajit G Pillai, Hafeni Mthoko, Shaimaa Lazem, and Marly Samuel. 2024. Embracing Ubuntu: Mapping Communal Ecologies in Participatory Design for HCI Practice. In *Proceedings of the Participatory Design Conference 2024: Exploratory Papers and Workshops-Volume 2*. 225–229.

Paresha Farastu, Nicholas Mattei, and Robin Burke. 2022. Who Pays? Personalization, Bossiness and the Cost of Fairness. https://doi.org/10.48550/ARXIV.2209.04043

Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. What is fair? Exploring the artists' perspective on the fairness of music streaming platforms. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II*. Springer, Cham, Switzerland, 562–584. https://doi.org/10.1007/978-3-030-85616-8_33

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 144–152.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 2 (2002), 12.

Archon Fung. 2003. Survey article: Recipes for public spheres: Eight institutional design choices and their consequences. *Journal of political philosophy* 11, 3 (2003), 338–367.

Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400.

Iason Gabriel. 2022. Toward a theory of justice for artificial intelligence. *Daedalus* 151, 2 (2022), 218–231. https://doi.org/10.1162/daed_a_01911

Susan Gasson. 2003. Human-centered vs. user-centered approaches to information system design. *Journal of Information Technology Theory and Application (JITTA)* 5, 2 (2003), 5.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

Tarleton Gillespie. 2022. Do not recommend? Reduction as a form of content moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

Nahid Golafshani. 2003. Understanding reliability and validity in qualitative research. *The qualitative report* 8, 4 (2003), 597–607.

Susan T Gooden. 2015. *Race and social equity: A nervous area of government*. Routledge.

Charles Goodhart. 1975. Problems of monetary management: the UK experience in papers in monetary economics. *Monetary Economics* 1 (1975).

Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

Royston Greenwood, Amalia Magán Díaz, Stan Xiao Li, and José Céspedes Lorente. 2010. The multiplicity of institutional logics and the heterogeneity of organizational responses. *Organization Science* 21, 2 (March–April 2010), 521–539. `https://doi.org/10.1287/orsc.1090.0453`

Tricia A Griffin, Brian Patrick Green, and Jos VM Welie. 2023. The ethical agency of AI developers. *AI and Ethics* (2023), 1–10.

James Grimmelmann. 2010. Some skepticism about search neutrality. *The next digital decade: Essays on the future of the Internet* (2010), 435.

Arthur Gwagwa, Emre Kazim, and Airlie Hilliard. 2022. The role of the African value of Ubuntu in global AI inclusion discourse: A normative ethics perspective. *Patterns* 3, 4 (8 April 2022), 100462. `https://doi.org/10.1016/j.patter.2022.100462`

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. " Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–31.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research, Vol. 80. PMLR, New York, NY, USA, 1929–1938. `https://proceedings.mlr.press/v80/hashimoto18a.html`

Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research, Vol. 80. PMLR, New York, NY, USA, 1939–1948. `https://proceedings.mlr.press/v80/hebert-johnson18a.html`

Deborah Hellman. 2008. *When is discrimination wrong?* Harvard University Press.

Bernie Hogan. 2015. *From invisible algorithms to interactive affordances: Data after the ideology of machine learning.* Springer.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–16.

HRW. 2018. *" Only Men Need Apply": Gender Discrimination in Job Advertisements in China.* Human Rights Watch.

Rosalind Hursthouse. 2007. Virtue ethics. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Palo Alto, CA, USA. `https://plato.stanford.edu/archives/fall2007/entries/ethics-virtue/`

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency.* 49–58.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 375–385.

Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How different groups prioritize ethical values for responsible AI. In *proceedings of the 2022 ACM conference on fairness, accountability, and transparency.* 310–323.

Dietmar Jannach and Christine Bauer. 2020. Escaping the mcnamara fallacy: towards more impactful recommender systems research. *Ai Magazine* 41, 4 (2020), 79–95.

Christina Jenq, Jessica Pan, and Walter Theseira. 2015. Beauty, weight, and skin color in charitable giving. *Journal of Economic Behavior & Organization* 119 (Nov. 2015), 234–253. `https://doi.org/10.1016/j.jebo.2015.06.004`

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 383–390.

Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication.* IEEE, 1–6.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th international conference on data mining workshops.* IEEE, 643–650.

Stamatis Karnouskos. 2021. The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cognition, Technology & Work* 23, 4 (Nov. 2021), 659–667. `https://doi.org/10.1007/s10111-020-00649-6`

Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*. 690–700.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arXiv:1711.05144 [cs.LG]

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM Press, New York, NY, USA, 100–109. https://doi.org/10.1145/3287560.3287592

Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM Press, New York, NY, USA, Article alt06, 11 pages. https://doi.org/10.1145/3290607.3310433

Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I Hong. 2022. " Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–37.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

Cory Knobel and Geoffrey C Bowker. 2011. Values in design. *Commun. ACM* 54, 7 (2011), 26–28.

Ramaravind Kommiya Mothilal, Shion Guha, and Syed Ishtiaque Ahmed. 2024. Towards a non-ideal methodological framework for responsible ML. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

Joseph Konstan and Loren Terveen. 2021. Human-centered recommender systems: Origins, advances, challenges, and opportunities. *AI Magazine* 42, 3 (2021), 31–42.

Melvin Kranzberg. 1986. Technology and history:" Kranzberg's laws". *Technology and culture* 27, 3 (1986), 544–560.

Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring group advantage: A comparative study of fair ranking metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 674–682.

Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443

Alexander Larry and Michael Moore. 2016. Deontological ethics. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, USA. https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/

Manzoor Laskar. 2013. Summary of social contract theory by Hobbes, Locke and Rousseau. SSRN. https://doi.org/10.2139/ssrn.2410525

Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. 2020. Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. *arXiv preprint arXiv:2003.07680* (2020).

Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszar, and Rumman Chowdhury. 2022. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *arXiv preprint arXiv:2202.01615* (2022).

Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019a. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019b. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2020. From fairness metrics to key ethics indicators (keis): a context-aware approach to algorithmic ethics in an unequal society. *Available at SSRN* (2020).

Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

Jennifer C Lena, Omar Lizardo, Terence E McDonnell, Ann Mische, Iddo Tavory, Frederick F VE Wherry, Christopher A Bail, and Margaret Frye. 2019. *Measuring culture*. Columbia University Press.

Yung-Ming Li, Jyh-Hwa Liou, and Yi-Wen Li. 2020. A social recommendation approach for reward-based crowdfunding campaigns. *Information & Management* 57, 7 (Nov. 2020), 103246. https://doi.org/10.1016/j.im.2019.103246

E Allan Lind and Tom R Tyler. 1988. *The social psychology of procedural justice*. Springer Science & Business Media.

Charles Lindblom. 2018. The science of "muddling through". In *Classic readings in urban planning*. Routledge, 31–40.

Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM conference on recommender systems*. 467–471.

Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1544–1558.

Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th international conference on World Wide Web companion*. 626–633.

Merriam-Webster. 2002. Merriam-webster. *On-line at http://www. mw. com/home. htm* 8, 2 (2002).

Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.

Thaddeus Metz. 2011. Ubuntu as a moral theory and human rights in South Africa. *African human rights law journal* 11, 2 (2011), 532–559.

Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 714–722.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437* (2021).

David Mullins. 2006. Competing institutional logics? Local accountability and scale and efficiency in an expanding non-profit housing sector. *Public Policy and Administration* 21, 3 (Sept. 2006), 6–24. https://doi.org/10.1177/0952076706021003

Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp., new york, usa*, Vol. 1170. 3.

Devesh Narayanan, Mahak Nagpal, Jack McGuire, Shane Schweitzer, and David De Cremer. 2024. Fairness perceptions of artificial intelligence: A review and path forward. *International Journal of Human–Computer Interaction* 40, 1 (2024), 4–23.

Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Aviv Ovadya. 2021. Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure — belfercenter.org. Available at: https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-\partisan-pressure. [Accessed 22-12-2024].

Aviv Ovadya. 2023. Deliberative Polls, Citizen Assemblies, and an Online Deliberation Platform — reimagine.aviv.me. https://reimagine.aviv.me/p/deliberative-poll-vs-citizen-assembly-meta-pilot. [Accessed 22-12-2024].

Anne-Claire Pache and Filipe Santos. 2013. Embedded in hybrid contexts: How individuals in organizations respond to competing institutional logics. In *Institutional logics in action, part B*, Michael Lounsbury and Eva Boxenbaum (Eds.). Emerald Group Publishing Limited, Bingley, UK, 3–35.

Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*. ACM Press, New York, NY, USA, 11–18. https://doi.org/10.1145/1454008.1454012

Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. *arXiv preprint arXiv:2201.12662* (2022).

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54, 1 (2010), 209–228.

Amifa Raj and Michael D Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–736.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 7 (April 2021), 23 pages. `https://doi.org/10.1145/3449081`

John Rawls. 1991. Justice as fairness: Political not metaphysical. *Equality and Liberty: Analyzing Rawls and Nozick* (1991), 145–173.

John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.

John Rawls. 2020. *A theory of justice: Revised edition*. Harvard university press.

Trish Reay and C Robert Hinings. 2009. Managing the rivalry of competing institutional logics. *Organization Studies* 30, 6 (June 2009), 629–652. `https://doi.org/10.1177/0170840609104`

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

Rwamahe Rutakumwa, Joseph Okello Mugisha, Sarah Bernays, Elizabeth Kabunga, Grace Tumwekwase, Martin Mbonye, and Janet Seeley. 2020. Conducting in-depth interviews with and without voice recorders: a comparative analysis. *Qualitative Research* 20, 5 (2020), 565–581.

Mark Ryan, Eleni Christodoulou, Josephina Antoniou, and Kalypso Iordanou. 2024. An AI ethics 'David and Goliath': value conflicts between large tech companies and their employees. *AI & SOCIETY* 39, 2 (2024), 557–572.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–26.

Steve Sawyer and Mohammad Hossein Jarrahi. 2014. Sociotechnical approaches to the study of information systems. In *Computing Handbook, Third Edition: Information Systems and Information Technology*, Heikki Topi (Ed.). CRC Press, Boca Raton, FL, 5–1.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 99–106. `https://doi.org/10.1145/3306618.3314248`

Frederick Schauer. 2018. On treating unlike cases alike. SSRN. `https://ssrn.com/abstract=3183939`

Morgan Klaus Scheuerman. 2024. In the Walled Garden: Challenges and Opportunities for Research on the Practices of the AI Tech Industry. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 456–466.

Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2021. Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine* 40, 2 (2021), 81–94.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

Amartya Sen. 2009. *The idea of justice*. Penguin Books London.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

Herbert A Simon. 1978. Rationality as process and as product of thought. *The American economic review* 68, 2 (1978), 1–16.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring user opinions of fairness in recommender systems. *arXiv preprint arXiv:2003.06461* (2020).

Jessie J Smith and Lex Beattie. 2022. RecSys Fairness Metrics: Many to Use But Which One To Choose? *arXiv preprint arXiv:2209.04011* (2022).

Jessie J. Smith, Lex Beattie, and Henriette Cramer. 2023a. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023 ((WWW '23))*. ACM Press, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583204

Jessie J Smith, Lex Beattie, and Henriette Cramer. 2023b. Scoping fairness objectives and identifying fairness metrics for recommender systems: The practitioners' perspective. In *Proceedings of the ACM Web Conference 2023*. 3648–3659.

Jessie J Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. 2023c. The many faces of fairness: Exploring the institutional logics of multistakeholder microlending recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1652–1663.

Jessie J Smith, Blakeley H Payne, Shamika Klassen, Dylan Thomas Doyle, and Casey Fiesler. 2023d. Incorporating ethics in computing courses: Barriers, support, and perspectives from educators. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 367–373.

Jessie J Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend Me? Designing Fairness Metrics with Providers. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2389–2399.

Janet Smithson. 2008. Focus groups. *The Sage handbook of social research methods* (2008), 357–370.

Nasim Sonboli, Robin Burke, Zijun Liu, and Masoud Mansoury. 2020a. Fairness-aware Recommendation with librec-auto. In *Fourteenth ACM Conference on Recommender Systems*. 594–596.

Nasim Sonboli, Robin Burke, Nicholas Mattei, Farzad Eskandanian, and Tian Gao. 2020b. "And the winner is...": Dynamic lotteries for multi-group fairness-aware recommendation. arXiv:2009.02590 [cs.IR]

Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020c. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 239–247.

Sonboli, Nasim and Smith, Jessie J., Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2459–2468.

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM Press, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372

Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, et al. 2022. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv preprint arXiv:2207.10192* (2022).

Rachel Thomas and David Uminsky. 2020. The problem with metrics is a fundamental problem for AI. *arXiv preprint arXiv:2002.08512* (2020).

Patricia H. Thornton, William Ocasio, and Michael Lounsbury. 2012. *The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process*. Oxford University Press, Oxford, UK. https://doi.org/10.1093/acprof:oso/9780199601936.001.0001

Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287566

Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

Niels Van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170 (2023), 102954.

Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.

Sahil Verma, Ruoyuan Gao, and Chirag Shah. 2020. Facets of fairness in search and recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

Patricia Victor, Martine De Cock, and Chris Cornelis. 2011. Trust and recommendations. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer, Boston, MA, USA, 645–675. https://doi.org/10.1007/978-0-387-85820-3_20

Amy Voida, Lynn Dombrowski, Gillian R. Hayes, and Melissa Mazmanian. 2014. Shared values/conflicting logics: Working around e-government systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3583–3592. https://doi.org/10.1145/2556288.2556971

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.

Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. ACM Press, New York, NY, USA, 1791–1800. https://doi.org/10.1145/3447548.3467289

David Gray Widder, Laura Dabbish, James D Herbsleb, and Nikolas Martelaro. 2024. Power and Play: Investigating" License to Critique" in Teams' AI Ethics Discussions. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–23.

David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.

Wikipedia. 2023. Discrimination. https://en.wikipedia.org/wiki/Discrimination

Daricia Wilkinson, Michael Ekstrand, Janet A. Vertesi, and Alexandra Olteanu. 2023. Theories of Change in Responsible AI. In *CRAFT Session at The 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

C Xu and T Doshi. 2019. Fairness indicators: scalable infrastructure for fair ML system. *Mountain View (CA): Google (accessed 2020-01-27). https://ai. googleblog. com/2019/12/fairness-indicators-scalable. html* (2019).

Emre Yalcin and Alper Bilge. 2022. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management* 59, 6 (Nov. 2022), 103100. https://doi.org/10.1016/j.ipm.2022.103100

Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image cropping on twitter: fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday* (2024).

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. 2021. Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. ACM Press, New York, NY, USA, 2439–2449. https://doi.org/10.1145/3447548.3467376

Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM Press, New York, NY, USA, 1153–1162. https://doi.org/10.1145/3269206.3271795

John Zoshak and Kristin Dew. 2021. Beyond Kant and Bentham: How ethical theories are being used in artificial moral agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM Press, New York, NY, USA, Article 590, 15 pages. https://doi.org/10.1145/3411764.3445102

Literature Review

[Lit1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 199 (nov 2019), 30 pages. https://doi.org/10.1145/3359301

[Lit2] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 international conference on management of data*. 1259–1276.

[Lit3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[Lit4] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3209978.3210063

[Lit5] Anubrata Das and Matthew Lease. 2019. A conceptual framework for evaluating fairness in search. *arXiv preprint arXiv:1907.09328* (2019).

[Lit6] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).

[Lit7] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[Lit8] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*. PMLR, 172–186.

[Lit9] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

[Lit107] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[Lit11] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443

[Lit12] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021a. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866

[Lit13] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021b. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.

[Lit14] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.

[Lit15] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 626–633. https://doi.org/10.1145/3041021.3054197

[Lit16] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595

[Lit17] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[Lit210] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys*

'18). Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372

[Lit19] Saúl Vargas and Pablo Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender systems*. 145–152.

[Lit20] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[Lit21] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).

[Lit22] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[Lit23] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

[Lit237] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1153–1162.

Appendix

Study #2

Included in this section of the Appendix:

- The collaboration document used with content creator participants during focus groups in study #2.
- The focus group protocol for content creator participants from study #2.

Welcome to our collaboration document!

We will be referencing and filling out this document together throughout the duration of our focus group.

Focus Group Agenda

Time	**Activity**
10:00 - 10:10	Arrivals and Logistics
10:10 - 10:20	Introductions
10:20 - 10:45	Unfairness Activity
10:45 - 10:50	5 Minute Break
10:50 - 11:05	Fairness Goals and Definitions Activity
11:05 - 11:25	Fairness Metrics Activity
11:25 - 11:30	Closing Thoughts & Discussion

First page of the focus group study collaboration document.

Arrivals and Logistics

10:00 - 10:10

Thanks for being here! We're excited to get started.

While we wait for everyone to arrive at the focus groups session, please spend a few minutes filling out this survey.

LINK TO FORM

All of the questions in this survey are optional. We ask demographic questions so that we can report the demographics of our participants in the resulting research publication, to be transparent about the generalizability of our findings and to improve the scientific rigor of our study.

Once you finish filling out the form, you can spend a few moments familiarizing yourself with this document to see what is coming up next!

Helpful Definitions and Terms

Fairness: impartial and just treatment of individuals or groups of people.
Algorithm: a combination of rules and guidelines that a computer has to follow.
Recommender System: an algorithmic system that curates large quantities of items (such as content/videos/images) and recommends them to an end-user (e.g., TikTok's "For You" page).
Ranking Systems: the final part of a recommender system, this chooses a ranking for the top recommended items (e.g., it selects the top 5 items to recommend to you next).
Providers: people who provide items to be recommended (e.g., content creators).
Consumers: people who consume recommended items (e.g., Instagram users).
Metric: a quantitative estimate of something (e.g., a measurement of number of views or a measurement of video engagement).
Exposure: how much your content is shown to viewers / end-users.

Second page of the focus group study collaboration document.

Introductions

10:10 - 10:20

Take a few minutes to brainstorm the following questions and prepare to introduce yourself to the group. You can share as little or as much of the following as you'd like:

- Who are you?
- Which recommender systems do you use regularly as a *content creator*? Which platforms do you regularly post on?
- What kind of content do you create and share online?
- Share a story of something you have created and shared online before.
- What is your experience with the platforms that you post on?
- Think about when you first started posting on this platform vs. now, do you see a difference in your process and how you evaluate your content exposure?
- Do you generally understand how the recommendation algorithms work on these platforms?
- Have you ever tried to "hack" the algorithm?

Feel free to take notes here to reference as you introduce yourself:

Name/Alias	**Note Taking Space**
Jess	This is an example of where I'd put my brainstorming notes
Aishwarya	This is an example of where I'd put my brainstorming notes

Unfairness Activity

10:20 - 10:45

5 Minute Brainstorming Session.

Brainstorm the following questions on your own and write down your thoughts in the provided space below.

Keep in mind that content is exposed on platforms like TikTok/Instagram/YouTube through recommendation and ranking algorithms. These algorithms decide which content gets shown to prospective viewers.

Questions (think of as many examples as you can):

- Have you ever felt like algorithms on platforms such as Instagram, TikTok, YouTube or Twitch have treated you or other content creators ***unfairly***? If so, how?
- What about these experiences felt unfair?
- Have you ever felt like algorithms on platforms such as Instagram, TikTok, YouTube or Twitch have treated you or other content creators ***fairly***? If so, how?
- What about these experiences felt fair?

Take notes here:

Name/Alias	**Note Taking Space**
Jess	This is an example of where I'd put my brainstorming notes

Fairness Concerns List (*Do Not Fill This in Yet*)

- *Fairness Concern 1*
- *Fairness Concern 2*

Chosen Fairness Concern: (*Do Not Fill This in Yet*)

- What makes this specific concern "unfair"?
- What would it look like if "fairness" was accomplished in this scenario?
- What actions would need to be taken to make this experience "fair"? (e.g., algorithmically, platform design, governance)
- Are there any tradeoffs that might result from this action? Does making the platform "more fair" for some people make it "less fair" for others? Who? Why?

Fifth page of the focus group study collaboration document.

5 Minute Break

10:45 - 10:50

Take a quick break from the screen, grab some water or a snack, and come back ready to complete the final 2 activities!

Sixth page of the focus group study collaboration document.

Fairness Goals and Definitions Activity

10:50 - 11:05

Platform of Choice: (*Do Not Fill This in Yet*)

5 Minute Brainstorming Session.

Brainstorm the following questions on your own and write down your thoughts in the provided space below. Refer to the example below if you need guidance.

Questions (think of as many examples as you can):

- What might be some examples of "fairness goals" for this platform?
- Write down your own definition of "fairness" as it relates to this platform and those goals. This should be one sentence long and very concise.

Take notes here:

Name/Alias	Note Taking Space

Jess (Example)	• ***Fairness Goals:***
	  ◦ *Do not perpetuate racism*
	  ◦ *Every content creator should have an equal opportunity to be shown to prospective viewers*
	  ◦ *Do not allow hate speech*
	• ***Fairness Definition:*** *Our platform seeks to ensure that everyone has an equal opportunity for their content to gain exposure, as long as it does not perpetuate hate speech, violence, or criminal activity.*

Fairness Metrics Activity

11:05 - 11:25

Platform of Choice: (*Do Not Fill This in Yet*)

10 Minute Brainstorming Session.

Brainstorm the following questions on your own and write down your thoughts in the provided space below. Refer to the example below if you need guidance.

Questions (think of as many examples as you can):

- Come up with your own quantitative and/or qualitative **fairness metrics** for this platform to measure how "fair" or "unfair" the platform is. For each metric, answer the following questions:
  - What data needs to be collected to measure fairness in this way?
  - What user populations (demographics) might need to be compared against one another? Why?
  - How will you know if "fairness" has been achieved?
  - How will you know if the platform is still not "fair" enough?

Take notes here:

Name/Alias	Note Taking Space
Jess (Example)	***Metric #1: Racism Metric*** • ***Goal:*** *Measure if the platform is exposing content from creators of all races at the same proportion.* • ***Data****: Would need to collect data on the race of each content creator, and data about which content is being exposed over others.* • ***Populations****: Would clump content creators into groups based on their race, and then would compare each groups' item exposure.* • ***Fairness Achieved:*** *If no race is exposed more than 10% above other races, fairness is achieved. The algorithm should not make assumptions about someone's race, this should be self-selected.*

	• ***Unfairness Remains:*** *If any race is being exposed more than 10% above other races, the platform is still not fair enough.*

Ninth page of the focus group study collaboration document.

Closing Thoughts & Discussion

11:25 - 11:30

Thanks so much for participating!

Feel Free to leave your content creator handle below if you'd like to stay connected with others from this focus group session!

- **Instagram:** @jessiejsmith
-
-
-

If you have any questions about this research or want to know about the results of this study later on, you are always welcome to reach out to the research team:

Jessie.Smith-1@colorado.edu

Content Creator Focus Group Protocol

20 Minutes: Intro Activity

- **10 minutes:** Introduce the research project
  - Play background music…?
  - Ask everyone to fill out the form about their demographics, and their email that they'd like to receive their gift card from / type of gift card
  - Let everyone know the goal of the research study + what to expect: *In this study, we are trying to understand your experiences with algorithmic systems that recommend and rank your content and choose when to expose it to an audience. We are going to discuss how these systems can be designed to treat you more fairly, and we will brainstorm ways to measure how fair the system is.*
  - Go over the agenda/plan for the focus group
  - **Ask for consent to start the recording**
- **10 minutes:** Have everyone introduce themselves and their current relationship with recommender systems as a content creator. **They should each answer the following questions:**
  - Who are you?
  - Which recommender systems do you use regularly *as a content creator*? (e.g., which platforms do you post on?)
  - Share a story about something they have created and shared online
  - Other possible brainstorming ideas:
    - What is their experience with these platforms and their algorithms?
    - Do they generally understand how the algorithms work? Do they ever try to "hack" the algorithms? Etc..

---

25 Minutes: Unfairness Activity

- **(5 minutes) Have them do a solo brainstorm session where they try to answer the following questions in the shared document:**
  - Have you ever felt like algorithms on platforms such as Instagram, Tiktok, Youtube, or Twitch have treated you or other content creators unfairly? If so, how? Have they treated you fairly? If so, how?
  - What about this experience felt unfair?
  - *Tips:*
    - *Tell them to think of as many examples as they can (not just 1!) And that they can use others for their examples, it doesn't have to be stuff that has personally happened to them*

First page of the focus group study protocol.

        - *Tell them they can also think about news articles or scandals where unfairness happened and use those as examples if they can't think of any personal examples.*
- **(10 minutes) Group Discussion: Have them discuss the following things:**
    - Have participants share the examples that they came up with while brainstorming
    - Based off your personal experiences, in what ways might these algorithms treat content creators unfairly? Fairly?
    - What are the general fairness concerns that providers on these platforms have?
    - **Deliverable: Research team will write down a list of all the fairness concerns related to this recommendation domain on the shared document as people are sharing.** *Tips: if they are struggling to come up with ideas, encourage them to discuss the ways that this platform causes algorithmic harm, ways that it perpetuates stereotypes, ways that it allocates resources and opportunities unequally, etc.*

- **(10 minutes) Group Activity:** Collectively select 1 example from the fairness concerns list that was just created by the group. **Have the group discuss the following:**
    - What makes this specific concern "unfair"?
    - What would it look like if "fairness" was accomplished in this scenario?
    - What actions would need to be taken to make this experience "fair" (algorithmically, in terms of platform design, or just generally like governance)?
    - Are there any tradeoffs that might result from these actions (e.g., does making it "more fair" for some people make it "less fair" for others? Who? Why?)

[5 MINUTE BREAK]

---

15 Minutes: Fairness Goals and Definitions Activity

- **(5 minutes) Have everyone in the group collectively choose a single platform related to their domain (e.g., TikTok, YouTube, Instagram). Tell them do a solo brainstorm session where they try to explore the following questions:**
    - What might be some examples of "fairness goals" for this platform? (For example: "Not perpetuating racism" could be a fairness goal)
    - Write your own definition of "fairness" as it relates to this platform and those goals. This should be one sentence long and very concise.
- **(10 minute) Group Discussion *(NOT IN THE SHARED DOC)*:**
    - *Have everyone share their fairness definition, and any fairness goals they wrote, look at what others wrote in the document*
    - How does everyone feel about their fairness definitions and goals? What conflicts or tensions or alignments arose? What challenges did you face?

- **How would you ideally want practitioners of [insert platform name] to determine fairness goals and definitions like the ones you just came up with?**
  - **Who should be involved in this process?**
  - **What methods should be used, and what questions should be asked?**

---

20 Minutes: Fairness Metrics Activity

- **(10 minutes) Have everyone do a solo brainstorming session where they try to answer the following questions (for the specific platform that was chosen in the last activity):**
  - Come up with your own quantitative and/or qualitative "fairness metrics" for this platform to empirically measure how "fair" or "unfair" the platform is. This includes answering the following questions for each potential metric:
    - What data needs to be collected to measure fairness?
    - What user populations (user demographics that we have access to) need to be compared against one another? Why?
    - How will you know if "fairness" has been achieved? How will you know if the platform is still not "fair" enough?
  - *Tip: tell them they can look at the example and also at what others are typing for inspiration while they are brainstorming!*
- **(10 minute) Group Discussion. Have them discuss the following questions (*NOT IN THE SHARED DOC)*:**
  - Have a few people share their ideas for metrics and their experience creating them.
  - How does everyone feel about the fairness metrics that were presented here?
  - What conflicts or tensions or alignments or challenges arose?
  - **How would you ideally want practitioners of [insert platform name] to determine fairness metrics like the ones you just came up with?**
    - **Who should be involved in this process?**
    - **What methods should be used, and what questions should be asked?**

Conclusion

I will wrap up the focus groups with a quick 5 minute class discussion about some of the challenges that people faced, and their general experience / takeaways.

- Explain that payment is going to happen soon, and you will reach out via email to let them know how.
- Give them the option to share their handles and stay in touch with one another
- Explain to them what the next steps of this research are and if they have any questions about the research or the results they are welcome to reach out at any time.

Study #4

Included in this section of the Appendix:

- The decision tree framework that resulted from Study #4 for both providers and consumers.

Decision Tree Framework (Consumers)

***Publication Note*: This version of the Decision Tree is not meant as a final product. It is a draft tool for the research community to further explore metric options available and their constraints.**

Instructions

Begin with a specific harm or unfair impact use-case for your recommender, ranking, or search system, then follow the decision-tree to find your options for appropriate fairness metrics to measure that impact. Once you find your way to a "leaf" node labeled "METRIC CATEGORY," you can find more information about those metric(s) and their associated academic papers by referencing their Fairness Category in Table 2. You can zoom in/out and scroll on this page to see text more easily.

Please note that the metrics linked from this decision tree are presented as options for your fairness measurement based on what is commonly used in literature. They are not meant to be prescriptive nor all-encompassing and we recommend you discuss these choices with relevant stakeholders before deciding on your metric(s). There may be more than one appropriate metric category for your use-case!

Key

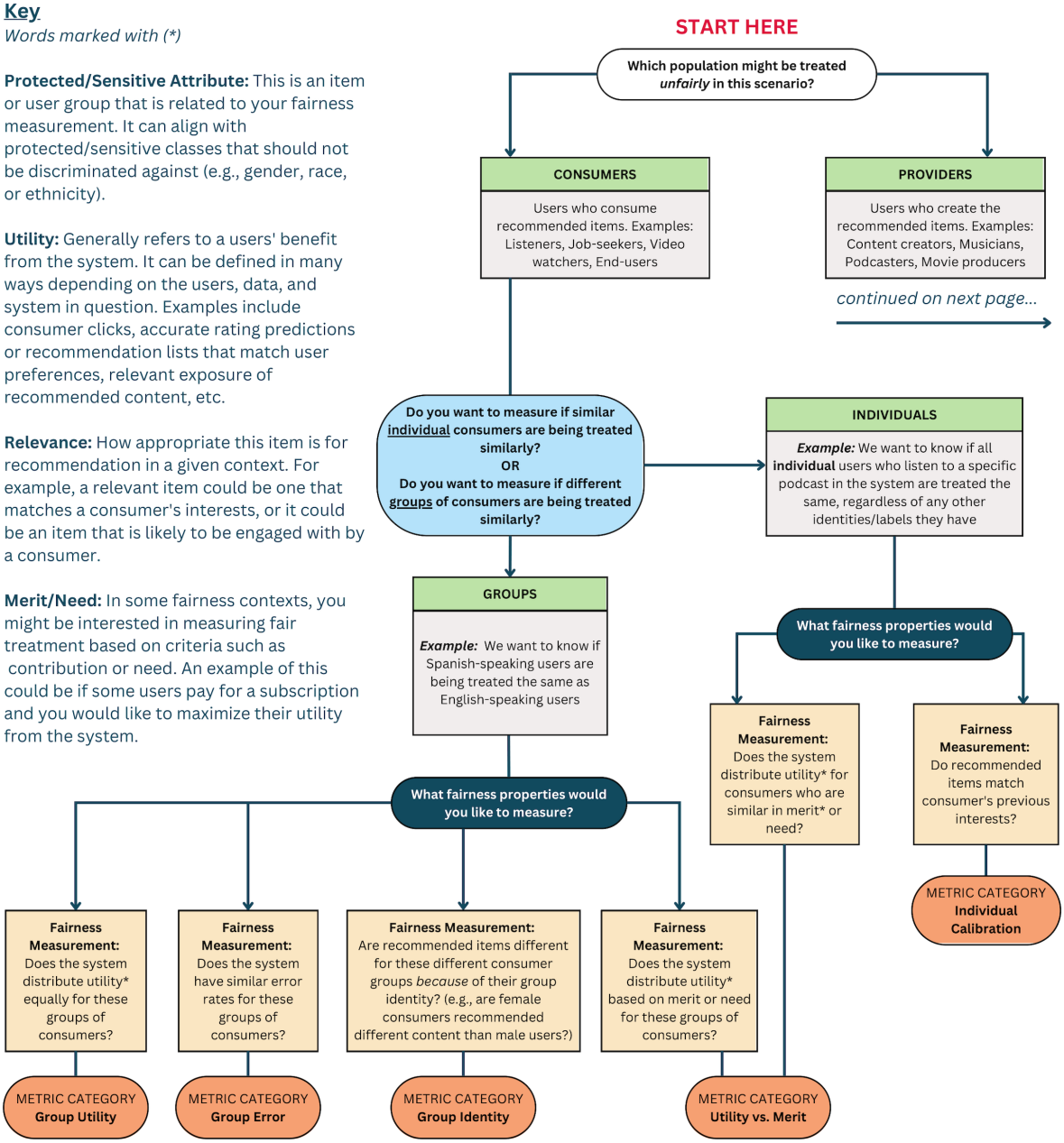
Words marked with (*)

Protected/Sensitive Attribute: This is an item or user group that is related to your fairness measurement. It can align with protected/sensitive classes that should not be discriminated against (e.g., gender, race, or ethnicity).

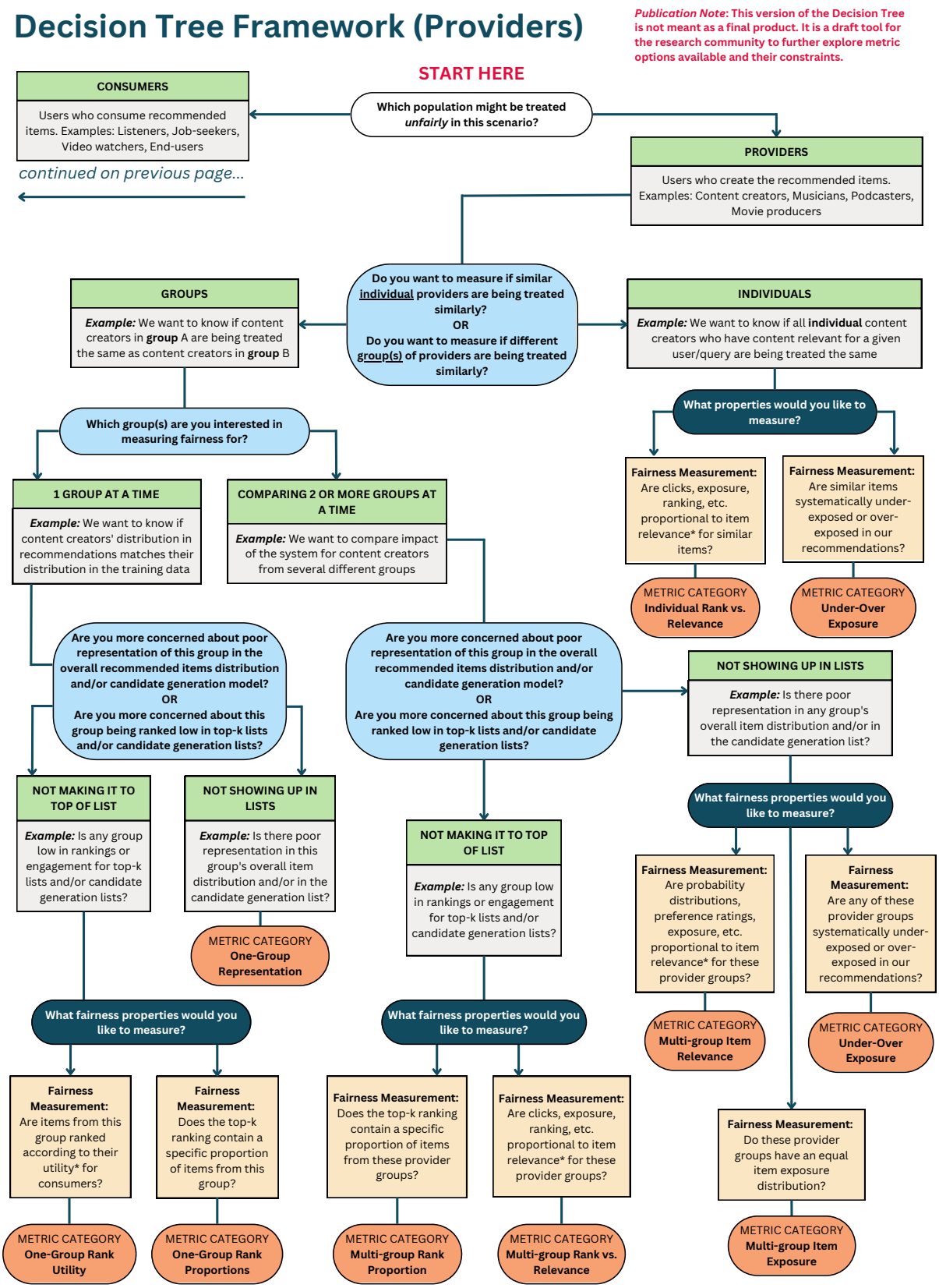
Utility: Generally refers to a users' benefit from the system. It can be defined in many ways depending on the users, data, and system in question. Examples include consumer clicks, accurate rating predictions or recommendation lists that match user preferences, relevant exposure of recommended content, etc.

Relevance: How appropriate this item is for recommendation in a given context. For example, a relevant item could be one that matches a consumer's interests, or it could be an item that is likely to be engaged with by a consumer.

Merit/Need: In some fairness contexts, you might be interested in measuring fair treatment based on criteria such as contribution or need. An example of this could be if some users pay for a subscription and you would like to maximize their utility from the system.



Decision Tree for Consumers from Study #4.



Decision Tree for Providers from Study #4.

Study #5

Included in this section of the Appendix:

- The protocol for interviews with ML fairness experts from Study #5.

Interview Protocol for Study #5

Preamble

Will say this out loud at the beginning of every interview

Welcome to the interview for uncovering best practices for fairness measurement in machine learning, using the case study of recommendation! During this interview, I will be asking you some questions about your experience with fairness in machine learning and recommendation systems, and we will discuss how you think fairness evaluations *should* be designed and conducted.

Before we begin, I want to confirm that you have seen and read the consent form for this study, and that you consent to participate in this research?

[If Yes] → Great! I am going to remind you that you have the option to not answer any questions that are asked during this study, and you are also allowed to leave the interview at any time.

May I begin recording this conversation?

Starter Questions (Part 1)

Goal: Warmup the participant and get them in the headspace of their evaluation experience and how they wish it had been better in the past.

Questions:

- What types of machine learning technologies and domains are you familiar with when it comes to fairness work?
- Do you have any hands-on experience with conducting fairness evaluations and/or operationalizing fairness?
- What challenges have you encountered when operationalizing or evaluating fairness in a machine learning system?
- What tools and resources have been helpful when scoping fairness goals, defining contextual fairness, or empirically measuring this definition (operationalizing fairness)?
    - What has been helpful about these tools?
    - What has not been helpful about these tools?
- Can you tell me about a time when you thought a fairness evaluation was done well? What aspects of this evaluation were done “well” in your opinion?
    - Potential follow-up / helper question: Tell me about a time when fairness evaluation was done poorly? How could this have been better?

First page of the interview protocol from study #5.

Experiences vs. Ideal vs. Challenges Questions (Part 2)

Goal: Learn how fairness evaluations are currently designed, how experts think they *should* be designed, which obstacles are preventing that ideal method, and which guidance would help practitioners achieve those ideal methods.

Script: *We are now going to spend the bulk of this interview going through each individual step of setting up a fairness evaluation. I'll begin by asking you about your experience with a given step of fairness measurement, and then I will ask you what you think would be the ideal best practices for that step. If we notice any differences between your experiences and your ideal best practices, we'll discuss why those differences might exist, and what might alleviate any challenges that are standing in the way.*

Experience *How have you or others on an evaluation team made concrete decisions about…*	**Ideal** *What do you think are the best practices for ML recommendation practitioners to…*
How have you identified potential **fairness concerns** or harms of the system you are evaluating?	How do you think practitioners *should* identify potential fairness concerns or harms of their systems?
How have you chosen what your **fairness goals** are?	How do you think practitioners *should* come up with fairness goals and considerations?
How have you **decided which user/stakeholder populations** to focus on with respect to fairness concerns / goals?	How do you think practitioners *should* select which user/stakeholder populations to focus their fairness goals on? ● Potential follow up: In the event that different stakeholders might have competing fairness goals, how should practitioners select who to prioritize in practice?
How have you **elicited fairness concerns or needs from stakeholders**? Follow up: If you haven't done this, why haven't you?	How do you think practitioners *should* elicit fairness preferences from real users and stakeholders of the system?
How have you **determined what your fairness definitions** are? (*Note that fairness definition here could be something like "equal opportunity", I am not speaking about the "mathematical fairness definition")*	How do you think practitioners *should* define fairness/unfairness based on those goals and considerations?

Second page of the interview protocol from study #5.

How have you **decided which fairness metric(s)** to use or to create for your evaluation?	How do you think practitioners *should* create or find metrics that capture their definitions of fairness? ● Potential Followup: What information would be most useful for practitioners to easily adopt existing metrics into their current evaluation pipeline? What kind of guidance might be necessary?
When you conduct fairness evaluations, do you typically start with a fairness goal/concern and then work towards a metric? Or start with a definition and works towards a metric? Or start with a metric and work backwards? What **order** are tasks done in?	How do you think practitioners *should* map their specific fairness goals to associated metrics and objectives that must be met by the system?
How have you determined what your **fairness / unfairness thresholds** are? (how do you determine when the results of an evaluation are fair enough or not fair enough?)	How *should* practitioners decide when the system is "fair enough" or "not fair enough" for deployment (e.g., through selecting fairness thresholds)? ● Potential Followup: Who should be making these decisions? Should they be monitored?
How have you prioritized **different/competing definitions** of fairness?	What do you think are the most ideal ways for practitioners to grapple with value tensions? What if they have multiple, competing fairness goals they want to achieve with their system? How should they prioritize these and how *should* this impact their evaluation? ● Potential Followup: What about guidance for dealing with value tradeoffs? How can we help practitioners recognize the different values at play and make decisions about which ones to prioritize?
Who would you say are typically involved in fairness evaluations you have been a part of? **Who are all of the decision makers**? Who conducts evaluations themselves?	Who do you think *should* be involved in fairness evaluation? ● <u>Potential Follow ups:</u> ● How should this fairness evaluation team be formed or built and how should they collaborate? ● Who should be the decision makers and who should be responsible for the evaluations themselves? ● Why X role? What kinds of expertise would X role help with ideally?
How have you assessed whether or not your fairness metric is actually measuring the thing they are trying to measure (**construct validity**)?	How do you think practitioners *should* assess whether or not their fairness metric is actually measuring the thing they are trying to measure?

Third page of the interview protocol from study #5.

Which parts of this evaluation process *are* usually **documented** in your experience?	Which parts of this evaluation process *should* be documented?

Language for Challenges / Barriers / Needs
[If there is a difference between their experience and their ideal best practice for a given step during the questioning above, ask this for each step!]
Did you notice any differences between your experiences and your description of ideal / best practices for fairness measurement?

- *If so, what obstacles might have prevented you or others from utilizing best practices in your fairness evaluations? Technical constraints? Organizational constraints?*
- *What would you need in order to do these ideal / best practices?*
- *What kinds of guidance would be most helpful for you and/or other practitioners to conduct robust fairness evaluations for recommender systems?*
    - *What kind of tooling is necessary to provide low-level, context-specific guidance and high-level, generalizable and scalable guidance?*
    - *What kinds of resources that aren't tools would be helpful for this? (e.g., worksheets, papers, checklists, mentorship…)*
    - *How can guidance be given that works for different datasets, systems, and contexts?*

Concluding Questions (Part 3):

Goal: If there is time, allow the participant to give any lingering comments / ideas that were not yet discussed.

Questions:

- Do you think there were any aspects of fairness evaluation that we did not discuss that are important to include?
- Any other thoughts or comments you'd like to add that are lingering from our conversation?
- Do you know anyone else who has expertise in recommendation fairness or ML fairness generally who might be a good fit for this study?
    - If so, would you mind sending them the recruitment information and letting them know they can reach out to me if they are interested in participating?

Fourth page of the interview protocol from study #5.